An Introduction To Educational Research

AN INTRODUCTION TO EDUCATIONAL RESEARCH

THE MACMILLAN COMPANY
NEW YORK · CHICAGO
DALLAS · ATLANTA · SAN FRANCISCO
LONDON · MANILA

IN CANADA
BRETT-MACMILLAN LTD.
GALT, ONTARIO

AN INTRODUCTION TO EDUCATIONAL RESEARCH

Robert M.W. Travers
Chairman, Department of
Educational Psychology
University of Utah

THE MACMILLAN COMPANY NEW YORK

First Printing

Library of Congress catalog card number: 58-9684

The Macmillan Company, New York
Brett-Macmillan Ltd., Galt, Ontario

Printed in the United States of America

Preface

This book is designed to serve two main purposes. First, it is planned for use in the training of educational research workers. Second, it provides an interpretation of the aims and methods of educational research for the large body of teachers, administrators, and students of education who use research results as a part of their daily routine but who are not and who are never likely to be research workers.

The research worker in education must have the support of the professional educator. If he wishes to study problems of classroom teaching, he must have the backing of both teachers and school administrators. If he wishes to study problems of teacher education, he must obtain the help and cooperation of the faculties of schools of education and teachers' colleges. The professional educator is likely to give support to research if he understands its purposes, but such understanding is often difficult to acquire because the research worker tends to live in a world of his own and to speak his own language. It is hoped that this volume will serve as a channel of communication between the research worker and those in education who are consumers of research results. Perhaps this volume may help to develop further the cooperative relationship that must exist between the

practitioner and the scientist in the field of education. The practical educator must come to understand why he cannot press the research worker for immediate and useful results and why he must often tolerate the researcher's indirect and often apparently up-in-the-clouds approach to problems of education. This book seeks to establish a mutual understanding between the scientist and the persons who actually conduct and administer educational enterprises.

The training of professional research workers is a matter of the greatest importance to the future development of education. The author feels that there is a real need for a book that helps the apprentice in the field of research to find his way around in the new world of ideas and methods he is entering. Schools of education are becoming more and more concerned with problems of preparing future generations of research workers. There is dissatisfaction with what has been done in the past, and an encouraging willingness to try out new approaches. This is a matter that has been discussed widely within the American Educational Research Association, where there has been concern over the fact that much educational research of the past few decades has been rather amateurish, largely because it has been undertaken by amateurs. This is not to say that there has been no progress in this matter, for there has been. Many schools of education have developed notable research departments, which are having a profound influence on the professional development of research. Nevertheless, an observer of the current scene cannot help being impressed with the fact that much of the current research activity is undertaken by persons who have devoted relatively little of their lives to this type of pursuit. This may be contrasted with research in other areas, where the typical research worker has devoted most of his life to the discovery of new principles. Just as the amateur has ceased to play a role in the development of knowledge of the physical sciences, so too must he cease to play a role in the development of systematic knowledge about educational processes. It is hoped that this book may play some small part in developing the new generation of professional educational research workers.

It is hoped, too, that the reader will realize that this book has been written in a spirit of humility. Very little is known about how man acquires a knowledge of his environment. Those who make major contributions to knowledge are not able to tell others just how

it is done. The fact that we have come to recognize our ignorance in this matter constitutes an important advance. When the writer attended graduate school a course was provided that supposedly taught the student how to do research, and there was little question on the part of either student or professor that this course achieved its set objectives. Such was the practice throughout the United States at that time, but most of those who were exposed to it have long since suspected that perhaps the skills of research were not so easily acquired after all.

This book is not an attempt to provide a quick and easy path to becoming a highly proficient educational research worker. It is realized that proficiency in research comes through years of direct experience with the process of inquiry and discovery. This work can help the beginner avoid some of the common pitfalls. Perhaps it can help him to obtain an overview of the current activities and methodologies in the field and acquire a critical attitude toward them. It is hoped that the reader will develop some understanding of the role of theory in a world of fact, and that he will come to demand that educational theories be stated with the same degree of precision required of theories in other disciplines. Perhaps through this book the student will acquire some humility toward the task of acquiring new knowledge and will come to realize that even the professional research worker with long years of experience is not able to produce at will. Every professional knows that there are long periods in his life when, for unaccountable reasons, he is totally unproductive. This book can perhaps help, and even encourage, but it offers no true and tried formula for producing original research.

The writer's opinion is that the book should be used in a course in which the student is given some opportunity to plan a research project of his own. The project does not necessarily have to be carried through to completion, but it should provide experience in research-planning, which is a vital and most neglected aspect of the training of the research worker. The student should actively attempt to gain experience in the various phases of research while he is studying this volume, even though he has no plans to become a research worker, for this is a sound way of learning. The instructor should help and encourage him to have such experiences. Even though the research activities undertaken in conjunction with this study may

be the only ones of his career, they may nevertheless give the student an understanding that he should have as a user of research results.

Reading about research should be encouraged by the instructor only insofar as it is helpful in enabling the student to profit from planned research experiences. For this reason, the references listed at the back of the volume serve primarily the purpose of providing sources of evidence to support statements made in the body of the book. They are not presented as a list of further readings for the student. This is an important point. Enough is presented here to guide the student through his early research activities.

The writer is indebted to his many colleagues and associates who have provided the extensive body of research knowledge from which he has drawn so freely. It would not be possible to list the names of the many persons from whom he has gleaned ideas and who have influenced his thinking, so this collective note of appreciation is offered. Another source of inspiration in the writing of this book has been the current emphasis in the American Educational Research Association on the professionalization of educational research. It is hoped that this book will make some small contribution to this effort which is already so well supported.

Thanks are due to my former associate Professor John Schmid, Jr., now at the University of Arkansas, who in detail read an early draft and offered many helpful suggestions. Professor H. Glenn Ludlow and Professor Finley Carpenter of the University of Michigan, the publishers' readers, offered invaluable advice and also some of the encouragement needed for reworking the manuscript to make it more useful to the student. Professor Carson McGuire of the University of Texas provided a particularly useful commentary on a much later draft. His forthright and indisputably sound criticisms resulted in the rewriting of several major sections. To my wife Norma I am indebted for her patient help over several years in preparing drafts of the manuscript.

Salt Lake City, Utah

ROBERT M.W. TRAVERS

Contents

Illustrations

Tables

AN INTRODUCTION TO EDUCATIONAL RESEARCH

Groundwork for Research | 1

The Traditional Formula for Educational Research

The older textbooks that discuss the methodology of educational research provide the graduate student of education with a simple recipe guaranteed to yield a doctoral dissertation of acceptable quality. The recipe usually runs more or less along these lines: First, the student is advised that he should find some hypothesis to test. Usually the explanation is that a hypothesis can be considered as the statement of a relationship between events, a relationship that is expected to be confirmed through the research that is to follow. Second, the student is expected to collect data that are relevant to the testing of his hypothesis. Third, statistical tests are applied to the data in order to determine whether they are to be considered consistent with the hypothesis. This simple recipe has been offered to generations of graduate students of education, and these same generations have been puzzled about why such a simple formula involves them in so many difficulties before a master's thesis or a doctoral dissertation is accepted by the faculty. The fact is that until relatively recently not enough was known about sound research procedures to provide more

useful advice, but during the last decades much that may at least help the student to identify his research difficulties has been learned, even if this body of knowledge does not lead directly to their solution.

A major defect in books that attempt to teach the would-be scientist how to set about his business arises from the fact that those who write on the subject are only rarely those who have direct knowledge of it. The reason for this is that much thought has been given to the subject by philosophers who have been immensely intrigued by it. The younger Mill was one such philosopher who wrote extensively on the nature of scientific investigation and yet never participated in any scientific inquiry. His views are of substantial interest to philosophers but are quite useless to the young scientist in helping him design explorations and experiments. There is a substantial difference between what philosophers say about the philosophy of science and what scientists say about their methods. Mill's works show a lack of familiarity with the difficulties that the scientist faces or the common causes of failure of scientific experimentation, because he never had any experiences that brought him into contact with these problems.

The difficulties that the scientist encounters are not easily identified by those who have not participated in such work. Indeed, even the mature scientist may have a limited conception of science because his own experience has been limited. The works of the philosopher who writes on science are more likely to be of interest to the mature scientist than they are to those embarking on their first scientific adventures.

B.F. Skinner (1956), famous experimental psychologist, has written of this matter, comparing his own personal experience of how discoveries are made with textbook accounts of this procedure. The following quotation well reflects the gap between the person who teaches *about* scientific methodology and the practicing scientist:

> But it is a mistake to identify scientific practice with the formalized constructions of statistics and scientific method. These disciplines have their place, but it does not coincide with the place of scientific research. They offer *a* method of science but not, as is so often implied, *the* method. As formal disciplines, they arose very late in the history of science, and most of the facts of science have been discovered without their aid. It takes a great deal of skill to fit Faraday with his wires and magnets into the picture which statistics gives us of scientific thinking. And most cur-

rent scientific practice would be equally refractory, especially in the important initial stages. It is no wonder that the laboratory scientist is puzzled and often dismayed when he discovers how his behavior has been reconstructed in the formal analyses of scientific method. He is likely to protest that this is not at all a fair representation of what he does.

The usual formula for conducting educational research is based on the idea that there is such a procedure as "the" scientific method. This is a notion that has had to be discarded, simply because scientific knowledge is arrived at by a variety of procedures and methods. Scientists vary greatly in this respect. A few prefer to follow the time-honored formula. Some, once they have even a vague idea of what they want to do, like to collect small samples of data and conduct rather rough-and-ready experiments; in this way, further cues are derived which may help to sharpen up a hypothesis. Still others may begin their explorations by reading widely in related fields, without too much concern about what they are looking for or what they may find. Some rely tremendously on personal hunches, while others reject anything that savors of intuition. *Techniques of arriving at knowledge, as they are manifested in the behavior of scientists, are highly personal and individualized. "The" simple scientific formula that all well-behaved scientists use simply does not exist.*

The reader may well ask at this point what a textbook on the methodology of educational research can possibly do for him, since it cannot teach him how to be an educational scientist. The answer is that, by discussing the methods some scientists use, a book of this kind can suggest useful techniques for arriving at scientific knowledge. It can also help the student avoid some of the pitfalls in which some of the more mature scientists have stumbled at times. It may prevent him from embarking on dead-end studies. Perhaps it may even add a little more polish to his finished product than would otherwise be the case.

An additional matter has been taken into consideration in the designing of a volume to help the graduate student in education. It is that the graduate student probably is working, so to speak, at the apprentice level rather than at the level of advanced discovery about which Skinner writes. At the student's level, advice is much more easily given than at higher levels.

Research activity may be considered in terms of levels. The highest

level involves the development of laws that apply to a wide range of phenomena. At the lower levels are found activities that are often no more than the application of familiar techniques to new situations. For example, it is possible to take one of the many techniques of factor analysis and apply it to a new field where it has never been used before. This is a type of research that could be handled by a well-designed machine. Research that could be undertaken almost entirely by any machine that man is now capable of building necessarily produces limited results, not the creative product of which the high-level scientist is capable.

The graduate student of education may reasonably be expected to undertake research at a somewhat higher level than that represented by the mere application of a complex technique to new data. His research should result in a generalization of at least limited applicability, but one would not expect him to make discoveries of the type to which the full-time researcher may aspire but rarely attain.

Educational Research: The Present Scene

The term *research* has come to be applied to such a wide range of activities within the field of education that it has ceased to have a single identifiable meaning. Within some school systems there are research departments that serve only the function of maintaining records of pupil enrollment and attendance and related data pertaining to the operation of the system. There are educational research organizations that devote their energies to the tabulation of data pertaining to such matters as the expenditures of the different states on education or the teacher-selection practices of different communities. Other educational research organizations administer tests, develop norms, prepare distributions, and engage in routine testing programs. Somewhat different are the activities of a few educational research institutes that conduct studies of variables related to and affecting the efficiency of learning, or studies of problems related to the development of personality. Such activities attempt to perform a function that goes far beyond that of data-gathering, because the data are collected for the purpose of deriving scientific generalizations that can be applied to the solution of a wide range of problems. This last meaning of the term *scientific research* is the sense in which it will be employed throughout this work.

From the point of view of this volume, educational research is that activity which is directed toward the development of a science of behavior in educational situations. The ultimate aim of such a science is to provide knowledge that will permit the educator to achieve his goals by the most effective methods. Presumably this will be accomplished by manipulating the child's environment so that it is as favorable as possible for fostering the desired direction of development. Since children differ in their abilities and other attributes, what is a favorable environment for learning in the case of one child may not be favorable for another. Individual differences enormously complicate the task that the educational research specialist sets for himself.

In order to clarify this point, consider some of the areas of education where a mature science of behavior in educational situations could be applied. It could be applied to the selection and training of teachers, to ensure that their behavior in the classroom was as effective as possible in promoting specific kinds of pupil change. It could be applied to the design of textbooks and other learning aids, to ensure effective use. It could be applied to the design of classrooms, not only to provide good physical conditions for personal comfort but to ensure that the social organization of the class was optimum. It could be applied to helping the pupil make long-range plans by helping him in his forecasts of what he could and could not accomplish. It could provide the principal with a sound basis for organizing the faculty and could guide him in providing conditions that permitted the teachers to develop to their full potentiality as professional persons. Thus one can continue. There is not a single phase of the educational process that a mature science of behavior could not render more effective.

Research and Value Judgments

A few contemporary philosophers have expounded the view that the role of research in education is minor, and that the central problems of education are problems of moral and ethical judgment. Once the major problems have been solved, these men claim, then research can play some small role in the solution of local problems. The following quotation from Mortimer Adler (1939) presents this point of view with some vigor:

> The basic problems of education are normative. This means, positively, that they are problems in moral and political philosophy; and, negatively, that they cannot, they have not and never will be solved by the methods of empirical science, by what is called educational research. . . . Neither the facts nor the generalizations [of science] can by themselves answer questions about what *should* be done in education . . . ultimate questions . . . are all moral. They cannot be answered by science. . . . [Ideals are not] relative and subjective, culturally determined or matters of individual opinion. . . . The major problems of education—whether in relation to the individual or to the state . . . have already been solved, for their solution does not depend on scientific research. Scientific research is relevant only in a minor connection, namely, the application of universal principles to local and contemporary circumstances. . . .*

In the judgment of the writer, this quotation presents a one-sided point of view. He prefers to believe that advances in thinking about the moral and ethical problems of education must surely go hand in hand with the acquisition of scientific knowledge about the behavior of persons in educational environments. What is the use of establishing lofty goals for education if it is later found that human beings are physically incapable of achieving these goals? What point is there in establishing standards of conduct that should be the mark of the educated person unless it is known just what are the heights that the human being can expect to achieve in this respect? It is clearly possible to propose solutions to what Adler believes to be the major problems of education, and then to find that they apply only to an imaginary universe. The philosopher and the scientist must surely work closely together on educational problems.

There is no doubt in the writer's mind that educational research would be greatly accelerated in its usefulness if the researcher were more often sensitive to the central ethical and moral problems of education. There is little worth in developing research on problems unrelated to these central issues. This does not mean, of course, that the researcher should try a head-on attack, for such a direct approach is often out of the question because of the unavailability of techniques. He can, however, work at points along the fringe of these problems where available techniques seem to apply, or at times he can seek to develop new techniques that are appropriate to particular aspects of

* From "Liberalism and Liberal Education," by Mortimer J. Adler, *Educational Record*, XX (1939), pp. 422–423. Reprinted by permission.

the problem. At all times he should remain aware of the nature of the central problem, even though he is working on the fringe.

Relation of Educational Research to Research in the Social Sciences

A science of behavior in educational situations of course should draw heavily on what has been learned in related fields. Psychology forms an important part of the background for educational research, but the other behavioral sciences also provide pertinent material. What has been learned about economic behavior is essential knowledge for those who would study the influence of economic factors on pupil behavior, teacher behavior, teacher selection, peripheral matters such as guidance services, and other phases of the educational process. The belief of the present writer is that educational research represents more than the mere application of the methods and theories of the related social sciences to educational problems. If it is ever to develop beyond the stage of an information-gathering activity, research in education must develop its own body of concepts, theories, and principles uniquely adapted to the ordering and prediction of events within the educational sphere.

The need for a framework to guide research is seen when it is understood that theories of behavior have been developed largely in special contexts. Theory of personality is derived largely from a clinical context. Learning theory has evolved largely in the animal behavior laboratory. Psychometric theory finds its background in industrial psychology. In each of these instances, a body of theory has been evolved that is uniquely adapted to the solution of problems in the area in which it has been developed. A misfortune of educational psychology is that those who have attempted to develop a science of behavior in classrooms have more often borrowed ill-adapted theories from other fields than they have developed ones that are useful in the study of educational phenomena. One might doubt, for example, whether a theory derived from the psychiatric study of seriously disturbed patients would be particularly useful in the study of day-to-day classroom events. The success that psychologists have achieved in developing theories in other applied areas suggests that the same success could be achieved in the development of theories of educational behavior.

There is much to be said for the view that most sciences have

been started by the discovery of principles that applied to a quite limited range of events, rather than universally applicable principles. For example, scientists had established the general nature of the laws of the motion of falling bodies and the laws of the motion of the planets around the sun long before Newton demonstrated that both of these phenomena could be understood in terms of more general principles, which are now known as Newton's laws of motion. Newton would undoubtedly not have arrived at his laws had not his predecessors provided him with a wealth of discoveries of limited significance. In recent times, Einstein has succeeded in integrating theories pertaining to motion, electromagnetic phenomena, and astronomical phenomena within a larger, more refined, and more comprehensive system. The point of this discussion is that the comprehensive theory is characteristic of a very mature science. In the early stages of the development of a science, the most profitable procedure seems to be to develop a set of theories each of which applies to a rather limited field. Attempts to develop comprehensive theories in the early phases of the growth of a science have been, in the past, notorious failures.

The reader has the right to ask at this point, "What does all this have to do with educational research?" The answer is that until recent times psychologists were concerned with the development of comprehensive theories of human behavior. The theoretical developments of Freud, although derived from limited clinical situations, illustrate well the comprehensive type of theory that was characteristic of psychology in the late nineteenth and early twentieth centuries. Freud's theory attempted to account for all forms of behavior. It included within its scope a theory of development, a theory of social as well as individual behavior, a theory of forgetting, a theory of creative behavior, and so forth. The weakness of psychoanalytic theory seems to have been in its comprehensiveness, for only in limited areas has it provided a useful basis for predicting and controlling behavior, specifically the behavior of clinical patients.

The same observation, generally speaking, is true of the other major systems of psychology that developed in the second and third decades of the present century. They have been, so to speak, universal systems designed to encompass all behavior; but experience in the application of such theories shows that their ambitiousness does not

bring with it a corresponding amount of success. Indeed, the success of such broad theories has been so limited that in recent times many psychologists have sought an alternative approach, which involves the development of theories that apply only to limited segments of behavior. For example, theories have been developed that cover such restricted phenomena as rote learning, discrimination learning, and speed of word recognition. Somewhat broader in scope are theories that attempt to organize the field of instrumental learning, or the function of drive in learning. Such theoretical developments of limited scope are widely regarded as the very essence of a modern science of psychology.

This trend in modern psychology has not been generally understood by educators, and misunderstandings have often led to its rejection as "impractical." Educators are heard saying that modern learning theories cannot be applied to classroom situations, and therefore the teacher must either fall back on some of the earlier psychological theories of learning or rely on common sense. What is not often appreciated is that these modern theories of learning are intended to apply to only the most limited domain, and those who develop them have no intention of applying them to educational problems. Indeed, any attempted application of this type would be regarded as unjustifiable generalization. Nevertheless, these same theories may provide concepts that can be used in the building of a theory of learning concerned primarily with the problems of classroom learning. The writer suggests that much too little effort has been made either to extract from current theories of learning those ideas which may have possible significance for education, or to use these extracted ideas as the foundation for new educational theories. Such new theories of classroom learning would be limited in their applications and might not be particularly useful for describing the learning and re-education that the clinician tries to produce in his patients. Thus, while educational research workers should watch the contributions made by scientists in related fields, they should also attempt to formulate theories that are specially adapted to the understanding of educational problems.

The need expressed here for a body of theories designed as a basis for educational research is not novel, if one is to judge from the history of educational thought. The most influential theory of

learning of the last century, at least insofar as education was concerned, was that propounded by Herbart—that human learning proceeded by adding to an "apperceptive mass." This theory found a basis for research and practice, and, although admittedly it was inadequately stated according to modern standards, its limited scope and educational context made it more useful than many modern theories. The later theory of learning developed by John Dewey also represents, not a comprehensive learning theory, but a theory of learning in school situations, and as such it has formed a basis for vast numbers of educational studies. Nevertheless, the Dewey formulation of learning in classroom situations should be considered a very rudimentary type of theory.

Just as the laws of falling bodies and the laws of the motions of the planets ultimately were integrated into a more comprehensive theory, so too may one assume that laws of learning used for predicting learning in classroom situations ultimately may be integrated with laws of learning derived in laboratory and other situations. In this way, more comprehensive theories of learning will slowly evolve.

Facts and Discoveries

Scientific research results in much more than the accumulation of items of information. The scientist cannot get along without information, but the mere accumulation of facts does not constitute scientific research. This often has not been properly understood by educators. Indeed, some agencies devoted to so-called research give their entire energies to the accumulation of information, and often such information is of doubtful value because it has been collected through questionnaires and other techniques of limited value. The product of the scientist is not a table of "facts" but a generalization. The generalizations that the scientist develops are usually called laws. These generalizations or laws must be such that they can be used to predict events. A generalization that applies only to past events is not a particularly useful one, although many of the generalizations derived from history are of this type.

The generalizations and laws of science are always based on considerable quantities of information, and sometimes this information is derived from daily experience. Newton's law of universal gravitation is a generalization derived from a great deal of daily experience as

well as from the specialized experience of the astronomer and the previous work of Johannes Kepler. The laws of thermodynamics also do much to bring order into innumerable daily experiences. Some generalizations are based only on the kind of data that the scientist collects. The discoveries of the great mathematician and astronomer Johannes Kepler would have been unthinkable if he had not acquired the voluminous data collected by his predecessor Tycho Brahe over a period of several decades.

Nevertheless, it is not enough for the researcher to have voluminous facts if he is to discover laws. Many students have arrived at graduate schools of education with files of data collected in their school systems and have found that somehow these vast quantities of facts could not be used to form the basis of a doctoral dissertation. Somehow, to most such individuals the data never suggest a problem that the data can be used to solve. Masses of data should not represent the starting places of research. The data for a major research should be collected only after the problem to be investigated has been well defined. Hypotheses that form the basis for a major research are not derived from masses of unorganized facts but from the available body of organized knowledge.

Theories and Laws

Implied in what has been said is the idea that theory has a place in educational research as in all other types of research. Although the assumption is accepted widely, the place of theory is not usually adequately recognized or identified. Too often theory is mentioned in terms of a gulf between theory and practice, and frequently with some ridicule concerning the theory side of this gulf. A distinction is made between the practical people who deal with facts (and who are alleged to "get things done") and the researcher who deals with theories. Such distinctions and discriminations serve only to produce confusion, because they are based on misunderstandings concerning the function and nature of theories.

Those who misunderstand do not recognize that all actions of a practical nature in educational situations are based to some extent on a theory of behavior. The teacher who attempts to enrich the curriculum with field trips and demonstrations is basing this action on a theory that learning is most efficient if the experiences provided

for learning occur in a variety of different milieus. The principal who institutes a series of staff conferences in order to install a new experimental curriculum is basing his action on a theory of social behavior insofar as his approach to the faculty is concerned, and a theory of learning insofar as the new experimental curriculum is concerned. The actions of practical people who operate educational programs nearly always are based on some kind of theory of behavior. In this respect only they differ from the researcher in that the researcher must state explicitly the nature of the theory underlying his work, while the practical educator does not have to do this.

Campbell (1952), who has written at some length on this phenomenon, points out that the practical man always seems to be willing to discuss his theories, which he has in abundance, but which differ from those of the scientist in both the way they are derived and the way they are used. The practical man's theories are formulations of what he has observed, but his observations always tend to refer to whatever events he *wanted* to see. A principal we know, who advocated teaching reading in kindergarten perhaps had arrived at this point of view by observing just one or two teachers who were particularly effective in teaching reading to a young group of very bright children. From that point on, he probably responded like most persons who have formulated or initiated a theory; he remembered only those subsequent instances that fitted his theory and forgot or disregarded the instances that did not.

One does not have to wander far in educational circles to find theories pertaining to every type and aspect of learning. There are advocates of socialized learning, individual problem-solving learning, rote learning, meaningful learning, learning by doing, and the rest. Many of these theories go back to such notable thinkers as Aristotle and Thomas Aquinas, and some are local in origin. Most, however, are based only on the type of observation that the scientist considers as merely a beginning for his activity while the lay person looks upon them as a sound basis for theorizing. The scientist takes up at the point where the layman rests his deliberation; but, as Coladarci (1954) has so neatly pointed out, "to relate research efforts, continuously, to theoretical considerations is not a disservice to 'practical' interests—they are mutually inclusive categories."

There are, nevertheless, certain real difficulties in bridging the gap

between the theories of the "practical" educator and those of the researcher. The theories of the former are couched in the language of the layman and are relatively easily communicated. Those of the researcher are stated in a technical language derived from the behavioral sciences and often are of a type that few scientists and far fewer laymen understand. Thus the practical educator, because he does not understand them, may often feel that the theories of the researcher have little application to actual educational problems. The ultimate interpretation of these theories into terms that the educator understands presents difficulties that have not yet been solved.

More often than not, educational policies and practices are based very largely on such popular theories. What is commonly referred to as educational theory is much more appropriately described as folklore than as science. The transition of theory from a folklore status to a scientific status is what the researcher in education aims to achieve.

Since the context of the discussion has been education, a contrast has been made between the theories of the practical educator and those of the scientist. A similar contrast could have been made between practical men in business, industry, and government and the scientific students in those fields.

The layman and the scientist also use theories in a different manner. The layman usually fails to recognize that a theory is only a tentative statement of a possible law, and he is likely to treat theories as if they were laws. All too often, the professional person acts like a layman. If he is a principal and believes that elderly teachers are more effective than young teachers, he is likely to use this credo as a basis for action and to favor elderly teachers in appointing new members to his faculty. He operates as if his theory were an established fact. The researcher does not do this when he is engaged in a study. In the pursuit of research, a theory is a beginning point, used to generate hypotheses that are later tested by experiment and inquiry.

A researcher, for example, has a theory that if a young child is given freedom, he tends to develop greater motivation to achieve and to succeed than if he is raised in an environment that places many restrictions upon him. He cannot test the theory as a whole in a single experiment or in a single study, but he does examine specific

aspects of it, which are called hypotheses. One such hypothesis might be that children whose parents greatly restrict their freedom to visit places outside of the home more frequently obtain relatively low scores on simple, routine tasks, such that the amount of work done is almost entirely a function of the subject's motivation. This is a specific testable hypothesis generated by the general theory on which it is based. Nevertheless, data collected in any particular experiment that substantiated the theory would only be what one might call circumstantial evidence. Evidence collected from various sources would be necessary before one could make any real statement concerning the use of the theory. The hypothesis might be supported in the study of some sample populations but not others, because several factors combine to explain behavior on routine tasks. In the behavioral sciences, the findings presented by various studies of hypotheses related to a single theory usually provide a picture with much greater ambiguity than do studies related to a physical theory.

When studies consistently support a particular theory, their results are customarily stated in the form of a set of generalizations or principles, which in turn permit the making of predictions. In the behavioral sciences, the generalizations and principles that thus far have been derived have only limited value in the making of predictions, and one must suppose that they are as yet based on rather poor theories and quite inferior evidence.

The experienced research worker who conducts a series of related studies, based on a common theory and designed to extend knowledge in a common field, is said to be conducting a programmatic type of research. Nearly all good research today is of that type. This presents a difficulty to the graduate student of education who conducts a single study that is likely to be both his first and his last. Consequently, a graduate student should seek to develop a study that is part of a continuing program with which he has become familiar through his reading or through personal contacts in the institution in which he is at work. Except in the rare instances, he should avoid investigating some isolated problem which has deep personal appeal and which he desires to investigate for that reason alone.

Constructs and Theories

In the behavioral sciences, one common practice is for the scientist to develop theories that postulate underlying mechanisms to account

for behavior as it is observed. In a sense, these ideas concerning underlying mechanisms may be considered to be products of the scientist's imagination, but they help him immensely in thinking about the phenomena that he is studying. These imaginary mechanisms are known as *constructs.* Sometimes they are referred to as *hypothetical constructs,* to indicate that they are not considered to be real objects or events. Most theories of behavior involve many constructs.

It is almost impossible to discuss behavior in terms of modern psychological theory without introducing constructs, and even much ordinary speech involves their use. We may speak of a person as having a liberal attitude, although we cannot observe his attitude directly; all we can ever observe is the result of this attitude as it is manifested in behavior. The attitude itself is a hypothetical construct introduced by the observer to "explain" consistency in behavior as it is seen. Abilities such as verbal ability, mechanical ability, and numerical ability are hypothetical constructs. The abilities themselves cannot be observed, for only behavior that *results from* these abilities is observable.

Unfortunately, in common speech there is a tendency to reify hypothetical constructs; that is to say, we tend to refer to them as if they were real and observable entities. An important part of the research worker's training is to learn to discriminate between hypothetical constructs and observable events, for this distinction is one that the scientist strictly observes.

Hypothetical constructs may be taken from many sources. First, there are those derived from neurology. Although relatively little is known about the functioning of the nervous system, something is known about the location of specific tracts and nuclei and about the transmission of impulses along these tracts. With this limited knowledge, it is possible to postulate the existence of certain mechanisms to account for behavior as it is observed. For example, the student who has read elementary textbooks on psychology usually is familiar with the diagrams of the supposed nerve mechanism that underlies the conditioned reflex. Now in actual fact, no person has ever directly observed such a mechanism. It is postulated on the basis of general knowledge of the nervous system. It is, in fact, a hypothetical construct introduced to account for behavior. A further example of neural constructs is seen in the type of associationist psychology with which the name of E.L. Thorndike is connected. In this sort of

theory, it has been common to think of the development of connections between stimuli and response as representing changes in the synapses, which are the areas of tissue that separate one nerve cell from another.

A second source of hypothetical construct is the scientist's own field of consciousness, or his *phenomenal field*, as it is called. Many constructs are derived from this source. Most of the constructs of the older systems of psychology are derived in this way from personal experience, but many modern psychologists doubt the usefulness of this procedure.

Nevertheless, the phenomenological theories do also have a strong group of supporters who work mainly in the clinical field. Carl Rogers, for instance, has exercised leadership within this group for many years, and a number of his students have also written extensively concerning this type of construct and the theories that result from its use. Despite the extensive literature they have produced, there is a striking absence of extensive experimentation based upon phenomenological formulations of human behavior, for experimentalists have usually favored other types of theories.

A third source of constructs is physics and mechanics. The psychologist Kurt Lewin, for example, drew extensively from these areas for his constructs. He argued that just as changes in the movement of a particle of matter could be considered as a result of forces acting upon it, so too could changes in the behavior of a person be considered the result of psychological forces. Lewin would draw diagrams to show the direction and the magnitude of the forces influencing behavior in a specific situation, just as the physicist draws diagrams to represent the forces acting on a particle. Other psychologists and educators have likened behavior to the motion of a particle within an electric field. So-called field theories of behavior based on this type of analogy have been widely used in educational thinking. Another analogy is the computing machine analogy, which is based on the fact that computing machines can be used to perform many functions that humans perform. The argument is that, since it is necessary to equip these machines with certain humanlike units (memory units) in order to perform the required functions, human beings may be considered to function *as if* they possessed analogous units. Usually those who use this analogy fully realize that it is only an analogy.

Definitive advice cannot be given at this time concerning the relative utility of the three main sources of constructs that have been discussed. The writer is under the impression that constructs derived from neurological and physical analogies are preferred by many contemporary theory-builders. Considerable mistrust is also evident of the type of theory that derives its constructs from the content of conscious personal experience. But the fact that a construct is derived from a particular area does not guarantee its utility. Each construct should satisfy at least one important condition, which must now be discussed.

In the development of constructs, the essential condition, as Hull (1943) has pointed out, is the avoidance of circularity of argument. This common defect in theory construction may be clarified by means of an example. In the study of problem-solving behavior, the custom in the past has been to "explain" such behavior in terms of a construct called intelligence. This procedure involves circular argument, for specific problem-solving behaviors are used as a basis for postulating an underlying ability referred to as intelligence, and then this underlying ability is used to explain the problem-solving behavior on the basis of which it was originally derived. In such a situation, the invention of a construct serves no useful purpose.

On the other hand, consider the case of a scientist working on a different problem. This scientist was concerned with the responses of subjects to a certain projective test, which he believed to be a measure of achievement motivation. In order to test this hypothesis, he was able to show that high scores on the test were related to the existence of childhood conditions judged to produce achievement motivation. The reader should note that in this latter case the construct of *achievement motivation* can be usefully introduced because it is related both to the conditions that produce it and the means through which it is measured. This construct involves no circularity of argument because it is firmly rooted both in the conditions through which it is produced and in the consequences observed in behavior. Such a construct is commonly referred to as one that is "tied down at both ends."

Philosophers such as Northrup (1948) have pointed out that in the early stages of a science the concepts introduced are developed on an intuitive basis, by which phrase is meant that they are derived from personal experience. In later stages, concepts are derived by

postulation; that is to say, they refer to hidden events that are inferred only rather indirectly from observed events. For example, the early stages of physics were characterized by concepts, such as that of force, that were derived from the direct and personal experience of tensions in muscles occurring when the body exerts a force on an object. In the later stages of physics we find concepts, such as that of the neutron, that are postulated on the basis of other events to which they are only remotely related.

Nearly all research on problems of behavior in educational situations is based on the more primitive of these two ways of deriving concepts. This is unfortunate, because the history of psychology shows clearly that research based on intuitive concepts derived from conscious experience is quite unproductive. The great advances of psychology have come when psychologists have postulated mechanisms and variables other than those which they could observe directly. Freud, for example, postulated a whole series of unconscious mechanisms that had no counterpart in conscious experience. All modern work on motivation is necessarily based on hypothetical constructs because individuals have no direct awareness of their own motives and indeed, according to clinicians, commonly misinterpret them. It is hard to find a field in which the study of behavior has advanced on an intuitive basis without the need for postulating hypothetical constructs.

A Conception of the Role of Theory in Research

Up to this point, the nature of theory has been considered only in the broadest terms. When the common man prefaces a statement with the remark "I have a theory that . . . ," he is saying that he believes he knows some law that will be useful to him. He may not be very sure that the law is a sound one or that it will apply in the specific instance, but he is inclined to believe that it does. Of course, he has not arrived at the law by any procedure that is acceptable to the scientist, and hence the scientist would question whether he really had arrived at a law. In contrast, when the scientist states that he knows a law that will serve the purpose of making a particular prediction or some other purpose, one can be sure that the law has been verified by determining whether it is capable of making the predictions that it is alleged that it can make. All laws have their

limitations and will work only within certain specified limits; and since these limits are rarely, if ever, precisely known, there is often question as to whether a law can be relied upon as a basis for a particular prediction. The theories that are used in educational research are usually represented by a series of generalizations about some aspect of education. These generalizations are based on information and are often substantiated by research, but they do not yet have the certainty, usefulness, or status of laws.

An example from a field outside of education may perhaps clarify this point more easily than one from education itself. From the chemical theory that burning represents a compounding of a substance with oxygen, it can be deduced that the products of burning must weigh more than the object that is ignited. Thus, in one of the classical experiments of chemistry, it was demonstrated that mercuric oxide resulting from the burning of mercury weighs more than the mercury from which it was derived, and this and similar evidence was collected to support the oxidation theory. Later, as chemistry grew to be a quantitative science, it became possible to predict from a more general theory just what would be the amount of oxygen that would combine with a particular substance.

In its early stage of development, the theory that burning is oxidation involved, like any other theory, one or more basic postulates from which deductions were made; or, to say the same thing but in different words, from which hypotheses were derived. In the simple chemical theory just considered there is a basic postulate from which deductions are made, and which is as follows:

Postulate: Burning is the combination of a substance with oxygen.

Deduction: The product of burning mercury will weigh more than the mercury that is burned.

Later, of course, examples were found of burning that did not involve the combination of a substance with oxygen, and the postulate had to be revised to include such instances. The theory that later evolved pertained to the heat changes produced when all types of chemical elements and compounds were combined.

An example from education may now be cited. The writer recently read a doctoral dissertation that was concerned with a problem re-

lated to the manifestation of aggressive behavior in children in classrooms. The theory on which the research was based depended on the following three generalizations:

1. Aggressive behavior on the part of the teacher results in aggressive behavior on the part of the pupil.
2. Restriction of movement in the classroom increases the amount of pupil aggression manifested.
3. Aggressive behavior on the part of the pupils tends to be manifested by the stronger toward the weaker.

These three generalizations about aggression in classrooms constitute a primitive theory. Although they are based upon some research, they could not be called laws; much additional research and verification would be needed before they could acquire that status. *Such statements are commonly referred to as postulates, which distinguishes them from laws. Postulates may be considered to be the forerunners of laws.* As more and more evidence concerning the validity of postulates is accumulated through research, they are modified where necessary. When found to be acceptable, they may finally be called laws.

The validity of postulates is examined by testing deductions from them. From the three postulates that have just been given, we may deduce (or "predict," if we wish to say it in that way) that children in classes with aggressive teachers will show more aggressive behavior in the home than those in classes with nonaggressive teachers. This deduction might also be called a hypothesis. If the hypothesis were tested in research, it might either confirm the validity of the theory or result in a modification of the postulates on which the theory was based.

In general, then, a theory is not a useful one unless deductions can be made from it. These deductions are specific consequences of the postulate or postulates to be tested. If the deductions are ambiguous, they cannot be used to test the validity of the theory; and unfortunately many theories of behavior provide deductions that are highly ambiguous. For example, there was a theory, often stated in articles in the early period of psychoanalysis, that implied that the psychological development of some individuals remained arrested at a so-called anal stage. Deductions from this postulate were that

such individuals would show in later years either excessive fastidiousness about cleanliness or, if this did not occur, a preoccupation with activities in which dirty materials were handled. Now this theory has almost everything wrong with it. First, it is not really possible to reduce it to a meaningful system of postulates. To do this, it would be necessary to specify the conditions that produce the condition known as anal fixation, so that we could write a postulate of the type, "X produces anal fixation." Then it would be necessary to write one or more postulates stating the general nature of the conditions that anal fixations do or do not produce. Finally, we would have to make deductions of the type, "When X is found in the background of the individual, we may later expect to find the symptom Y [Y is some specific consequence of an anal fixation] occurring more frequently than when X is absent." Obviously, it is simply not possible to test the validity of a theory when it is deduced that either a symptom or its reverse may occur in adult life as a result of a childhood event.

Throughout this discussion, the assumption has been made that the postulates or laws stated in the presentation of a theory are thoroughly understood by all. This is rarely the case, and hence it is almost always necessary to present a set of definitions of the terms that are used as a part of any theory. Sometimes variables that are mentioned in the postulates are defined in terms of the way in which they are measured. For example, a postulate of one theory includes the word "anxiety." In this theory, anxiety is measured and defined by means of the Taylor Manifest Anxiety Scale. The use of operations yielding an index or a measure is a common way of defining the variables mentioned in an educational theory.

Summary

1. There is no single well-tried formula that can be used for arriving at knowledge. Despite the attempts of philosophers to formulate simple procedures that will result in the production of scientific knowledge, such procedures do not exist at this time. Scientists differ greatly in the ways in which they arrive at scientific knowledge.

2. Educational research is considered in this volume as that activity which is directed toward the development of a science of behavior in educational situations. Thus it represents a branch of the behavioral sciences that has special implications for all phases of educational planning

and that would help the teacher to know what conditions to establish in the classroom in order to achieve particular results.

3. Educational research provides knowledge concerning educational objectives that *can* be achieved and indicates efficient ways of achieving them, but does not determine the ethical and moral values that education should foster.

4. While educational research may, and should, draw upon knowledge acquired in related sciences, it may nevertheless have its own unique characteristics.

5. A scientific body of knowledge acquired through educational research would not consist of a mere body of fact but would provide generalizations and laws that could be applied to the solution of a range of problems.

6. Theories are not just "ivory-tower" phenomena, but are developments of the greatest practical importance. A good theory also marks the point of departure from which a successful exploration of educational phenomena are made.

7. In the behavioral sciences, of which educational science is one, the research worker may build his theories from a number of different sources of materials. He may build them in terms of his knowledge of the anatomy and physiology of the human being. He may build them out of elements in the content of his own conscious experience. A third alternative is that he may borrow ideas from the physical sciences. Which type of material is used as a basis for theory-building is largely a matter of personal preference.

8. Scientific theories can be considered, under ideal circumstances, to consist of a series of laws related to one another. Most theories in education that are to be used as a basis of research cannot be considered to consist of a set of laws, but rather do they consist of a set of postulates that have much less validity than laws. In any case, a theory must include definitions of terms. Deductions from a set of postulates become the means of testing the adequacy of a theory. Such deductions are commonly referred to as hypotheses.

Conducting Research Within a Framework of Theory | 2

Current Educational Theory as a Basis for Research

Theories in their most rudimentary form are often no more than ways of looking at data. The behavioral sciences were at one time plagued by this type of theorizing, but perhaps one should take a less critical attitude toward such rudimentary theorizing and realize that it was the path that had to be trod before more useful and adequate theories could be developed. The reader will recognize that from Charcot and Freud up to modern times, clinicians have developed theories that were nothing more than ways of viewing the clinical process, for there was no satisfactory way in which the expected results of the theories could possibly be tested. For example, there is no possible way of testing the Freudian theory that nothing is ever forgotten, for it is quite impossible to conceive of an experiment in which it might be demonstrated conclusively that some event had actually been forgotten.

Any theory that has merit for scientific purposes should be such that one can conceive of evidence that might be inconsistent with it.

All scientific theories that have played a major role in the advance of knowledge have been of this type. The rudimentary type of theory we have considered here can be classed only as a kind of crutch to thought.

Theories may be stated with varying degrees of precision. Most educational theory is stated in an informal manner and in the language of everyday speech. Dewey's theory of problem-solving (1910) is an example of such an informal educational theory. One section of this theory, for example, implies that learning to identify accurately the problem to be solved is an essential aspect of learning to solve problems. This statement of a small segment of Dewey's problem-solving theory could be said to be an informal statement of one aspect of an educational theory. It attempts to state in a very general way what is believed to happen in one phase of the problem-solving process.

Such theories have often provided useful guides for action in the classroom, but they have had relatively little influence in providing guide lines for research. The writer is not aware of any study of consequence that has emerged from Dewey's theory of problem-solving. Theories need to be stated in somewhat more precise terms to be of value to the scientist, and while it may not be necessary or even possible to state educational theories in completely formal terms for satisfactory research to result, a greater degree of precision than that ordinarily found seems desirable.

The student's concept of some of the characteristics of a theory needed as a basis for teaching and research may be developed by examining a theory of some historical interest. Consider, for example, the theory of education of Dr. Montessori, which several decades ago aroused a great deal of interest. This theory takes as its primary postulate the statement that freedom of movement within the classroom is an essential condition for effective learning, and as a second postulate that certain objects have intrinsic value in stimulating the interest of young children. The latter objects, the apparatus of the Montessori system, were to be discovered by trial and observation. These postulates represent theory in a very rudimentary form. The deductions that can be made from them are only of the most direct type and involve no more than statements concerning the validity of the postulates themselves. Thus, one may "deduce"

from the first that children in a free-movement situation should learn more than children in a movement-restricted situation. A test of the validity of this "deduction" provides evidence of the validity of the postulate. There seems to be no way in which it is possible to make deductions from more than one postulate in a manner that links together the various elements in the theory. There is also only a limited rationale on which the postulates are based. They do not stand on a firm foundation of carefully collected data, but, rather, they are based on general observation.

The Montessori theory of education has merit as a theory, in that it stresses the manipulation of concrete conditions and therefore is closely tied to observable events. This simplifies the task of the teacher, who can manipulate the events and their related conditions. A similar desirable feature is not shared by most educational theories. They tend to be preoccupied with the personal experience of the pupil and the manipulation of events within his inner life. Such theories assume that by the provision of verbal materials and various visual and other cues it is possible to generate such personal experiences as "feelings of security," "understanding," and "thinking." Such theories rarely refer to variables that can be measured through some form of objective and recordable performance. The end products are hidden.

One of the handicaps presented by educational theories is that they are also often influenced by sentiment, rather than by a desire to manipulate conditions so as to make education as efficient as possible in the achievement of certain goals. This is reflected in the stress placed on the feelings of the teacher toward the pupil and similar traditional variables of personal experience. Theories based on sentiment often stress how the teacher should feel, as in primitive theories that emphasize that the teacher should love the pupils. Such theories tend to become somewhat mystical when they are probed to the point of answering the question concerning how the teacher's love of the pupil is functionally related to the learning process. Merely to state that it provides a favorable climate in which learning can flourish, just as certain regions have a climate in which grapefruit can flourish, does not provide a satisfactory answer to the question.

The last type of educational theory is particularly perplexing to a student of the behavioral sciences because it is based on postulates

that are inconsistent with a large body of data concerning both human and animal behavior. Despite this fact, it has been a widely held type of attempt to state a theory concerning the conditions necessary for effective education and was originally popularized in the writings of Jean Jacques Rousseau. It is a type of theory that is wholly useless to the scientist and of doubtful value to the educator. Statements of the theory never indicate to the teacher just what are the positive conditions that must be manipulated if the teacher is to achieve a particular objective in education. From the scientist's point of view, the theory does not generate a testable hypothesis.

The wide gulf that exists between scientific theories of behavior and educational theories does not have to exist. Ultimately, such theorizing may attain a certain unity. A theory that is clearly stated and provides a useful basis for action in the classroom also should provide a sound basis for research. The gulf that exists between educational theory and behavioral theory simply does not have to persist.

The Statement of a Theory as the Starting Point of a Research

From what has been said, one would infer that theories may vary greatly in their complexity. Some contain a single idea while others incorporate many. The single-idea theory, based on a single postulate, commonly has been used as a basis of master's theses. For example, one student started with the postulate, "The rate at which pupils learn to recognize words is a function of their ability to discriminate differences in shape." The student was careful to define what he meant by the ability to recognize words and the ability to discriminate differences in shape. He then proceeded to determine whether this was the case for a given group of words that a certain class of pupils was learning to recognize.

What would have been the advantage of having stated a more complex theory for the purpose of the research? It would have served the purpose of stating as clearly and as concisely as possible the state of knowledge in the area in which the student proposed to work. That is really the function of a theory that has been stated comprehensively. It gathers together the various ideas in the field, and when this has been done, it is possible to see more clearly what needs to be done and where are the major gaps in present knowledge. There is no absolute necessity to bring together the ideas in a field in order

to do research, but, since the research worker usually has made a careful review of the literature in the field, he might as well do this. If he does not organize the theory of his area of interest in this way, he is always in danger of basing his research on a single postulate that includes an idea or concept that really does not tie in with other already acquired knowledge.

Bergmann (1957) has very neatly stated this point when he asserts:

> A concept is significant if and only if it occurs, together with others, in statements of lawfulness which we have reason to believe are true. . . .
>
> Assume that somebody proposes a new concept, call it the C-coefficient. A person's C-coefficient is, by definition, the number obtained by multiplying his white blood count by his weight in ounces and dividing the product by the number of hairs on his legs. Clearly, it is not difficult to ascertain a person's C-coefficient. Equally clearly, the concept is not significant. Why, one may ask, are we so certain of this? After all, there could be laws in which it occurs. In principle this is so. Yet we are certain that there are none. To understand the reasons we have for this certainty, assume that the proponent of our new concept is a crank who expects to use it for the prediction of cancer, that is, he hopes to find a law that makes the incidence of cancer a function of the C-coefficient ("C" from "cancer"). Again, why do we call him a crank? The point is that we know a great deal about cancer, and that neither the C-coefficient itself nor the law our friend expects to find "fits" with existing laws and theories about cancer.*

What Bergmann is saying is that one should use as a basis for research only those concepts that are linked to other concepts. The isolated idea, however brilliant it may seem to the person who has generated it, has no real place in science, for the scientist attempts to build *relationships* among ideas. He builds upon the ideas of the past, though sometimes he may radically reorganize these ideas and see them in a new light. The related ideas that constitute a theory are the postulates of the theory. At this point we should begin to consider some examples of theories from the educational field, deriving our main example from the area of reading, where a very large amount of knowledge has been acquired.

* Reprinted with permission of the copyright owners, the Regents of the University of Wisconsin, from *The Philosophy of Science*, by Gustav Bergmann, Copyright © 1957, The University of Wisconsin Press.

An article describing a method of teaching reading in the very early stages of learning the skill provides an illustration of theory-construction and declarations from postulates. This article, which is widely considered to be one of the better pieces in the field, is based on a fairly definite theory of reading, although the theory is not very well stated. The nature of the theory may be represented by defining some of the key terms that were used by the author and then setting out the three major postulates that appear to form the very core of the theory. Definitions of the key terms and the three postulates as formulated by the present writer are set out in Table 1.

The next step is to determine whether the theory can be used as a basis for finding problems for research. In other words, what deductions that serve as hypotheses in a research program may be made from the theory? The writer then thought through some of the logical consequences of the theory and listed them as deductions. One can be sure that some of these deductions have already been tested, but probably at least one has not. Many other deductions also could have been listed.

To those who have been raised on Newtonian mechanics, this attempt to state an educational theory may appear to be pathetically inadequate. In the theory presented, the deductions *more* or *less* follow from the postulates, but they lack the tight logic of mathematical deductions. The theory lacks precision in the sense in which precision is found in the theory of physics. Yet it can hardly be denied that a theory stated in the fairly terse and organized form suggested in this chapter may be much superior for scientific purposes to a theory presented in a long and wandering article in which the postulates, deductions, and illustrations are all mixed together. Undoubtedly much could be done to improve the statement of the theory presented in the illustration, and the writer's attempt at theory-building probably does not do justice to the article from which it was derived.

TABLE 1. A Theory of the Early Stages of Learning Reading

Definitions

1. *Reading* is defined as a controlled form of talking in which the words that are said are controlled by the nature of the written symbols presented.

2. *A correct reading response* is defined as the act of saying the agreed-upon interpretation of the written symbol presented.

3. *Accuracy of response* to a word is defined as the percentage of attempts to say the word that are correct.

4. *The perception of learning to read as a goal* is evidenced by such behavior as the pupil asking the teacher for reading activities, participating voluntarily in reading activities, choosing reading activities rather than others.

Postulates

1. When reading is learned by means of the sequence: written word presentation, vocal response by the teacher, vocal response by the pupil, the frequency of occurrence of this sequence is related to the accuracy of response of the pupil. (Reader, note that this method of learning to read is commonly referred to as the "look-and-say method" and will be so referred to here.)

2. The effectiveness of the look-and-say method in generating correct reading responses in the pupil is related to the ability of the pupil to discriminate form and shape. Pupils must have a minimum of the latter ability if the method is to produce learning. Additional increments of the ability beyond the minimum result in increased rates of learning.

3. The effectiveness of the look-and-say method in producing correct reading responses is related to the extent to which the pupil perceives the learning of reading as a desirable goal and is motivated to achieve that goal.

Deductions

1. Measures of motivation to read will be correlated with accuracy of response in the early stages of reading in the case of those pupils who perceive reading as a desirable goal.

2. Failure to discriminate two words is a function of the similarity of the shape of the two words.

3. The look-and-say method produces greater accuracy of response when it is supplemented by procedures that emphasize the discrimination of the form of one word and the form of another than it does when such methods are not used.

Since any theory is likely to be a product of intensive study of the scientific knowledge available in an area, it is apt to appear to the person who has not undertaken such studies to be intensely tech-

nical and perhaps even incomprehensible. For this reason, some fairly simple examples from areas known to most readers are presented in the main body of the text. A few readers, however, may like to inspect a more sophisticated example of theory construction, even if it does involve a great deal of technical vocabulary. Therefore, an example of a theory which involves somewhat greater sophistication is included in an appendix. This example, drawn from Ammons (1954), obviously has been prepared after the most careful study and thought, and it provides a fairly complete example of a theory stated in terms of postulates, and one in which the vocabulary has been very carefully defined.

Level of Comprehensiveness of a Theory

Theories vary considerably in the extent to which they cover all of the factors that affect the events being studied. To explain this point, a fairly familiar example may be taken from the field of physics. Consider the problem of predicting the trajectory of a shell fired from a gun. By simple deductions from Newtonian physics, it is possible to approximate this prediction if the muzzle velocity of the shell is known and also the value of G, the gravitational constant. A more comprehensive theory would take into account the resistance of the atmosphere, the barometric pressure, and the direction and velocity of the wind. The more comprehensive the theory, the more precisely it is possible to predict what it is desired to predict, but this is true only insofar as it is possible to measure the variables that the theory includes.

In the development of a science of educational behavior, we can also build theories of varying degrees of comprehensiveness. We could, for example, build a theory of child behavior in the classroom that took into account a tremendous range of variables, including immediately preceding circumstances as well as circumstances far in the background of the child. Such a theory would have a high degree of comprehensiveness, but it would obviously not be completely comprehensive because it would not include many important variables that had not yet been recognized. Nevertheless, as a theory it would probably have little utility. It would involve too much that could not be measured at this time and too much that scientists will not, with any likelihood, be able to measure at any time in the fore-

seeable future. The main defect of such a theory is that it is *not* firmly rooted in current knowledge. A theory that is so rooted is likely to be relatively simple.

A simple theory that deals with relatively few major variables, which can be measured, can be a much more productive enterprise than one that deals with a larger number of variables, most of which cannot be measured. It is important in this respect, as in others, to prevent oneself from becoming mentally suffocated under a mass of detail. Relatively simple and incomplete theories have had a history of usefulness in the behavioral sciences and an unexpected degree of success in terms of what was anticipated fifty years ago. A good example of this is the theory that states that educational achievement is merely a function of a few variables such as the verbal factor, the numerical factor, the deductive reasoning factor, and so forth. Such a theory fails to take into account the fact that achievement is also a function of motivation and the presence or absence of numerous external conditions that favor or do not favor learning, for example the way in which the teacher interacts with the pupil. Despite these limitations, this simple type of theory has been the basis of a vast amount of productive research and has formed a basis for much that takes place within the general area of guidance. It has also become the foundation for the entire system of assigning men to training programs within the armed services. Its success has been nothing short of astounding.

Unfortunately, predictions made in terms of this simple type of theory have only moderate accuracy and leave approximately 50 per cent of the variability (variance) of the predicted variable unaccounted for in terms of the prediction variables. It is of considerable interest to observe in this connection that students of the behavioral sciences are not the only scientists who use effectively theories that have only, so to speak, 50 per cent efficiency. Many industrial manufacturing processes in the field of complex chemicals are based on traditional concepts of organic chemistry. On the basis of such chemical theory we may expect a given amount of component chemicals to yield 100 per cent of the chemical product that it is desired to manufacture. In practice, the manufacturing may in fact produce only 50 per cent of the expected yield. For reasons unknown at the present time, side reactions generate a great number of subsidiary

products, some of which are usable and some unusable. The theory is a valuable one in spite of its limitations, and it is certainly widely used as a guide in the development of industrial products.

The researcher wants to extend theories already existing so that they become more comprehensive, but not to a point where they involve many variables that cannot be measured at this time. Modest extensions may be extremely valuable, and these do not have to be grandiose to represent a substantial development over the previous state of the technique.

Formalizing a Theory

The reader undoubtedly has heard discussions of the need to formalize theories. The concept of a "formal theory" has been used with a great diversity of connotations, and often these two words are uttered as if they had some special magic power. The implication is that all one has to do is to state a theory in formal terms and great scientific achievements will result. But what is meant by a formal theory? The writer has a preference for the meaning used by Bergmann (1957), who regards the formal statement of a theory as one in which all of the words have been translated into abstract symbols such as are used by mathematicians. A theory stated in the form of a series of mathematical equations would be a formalized theory. The science of behavior in educational situations is not advanced to the point where a formal theory of this type is feasible. Perhaps we may ultimately aspire to the statement of educational theory in such terms. At the present time it is not known whether this is even possible.

Causal and Functional Relationships

It is stressed at various points in this volume that the educational research specialist should seek to establish organized systems of relationships among events and among variables. Such relationships can be stated without introducing the notion that some events cause other events. For example, from the data collected in the past concerning the motions of the planets in the solar system, positions of the planets at some future time can be predicted. The lawfulness of planetary behavior is neither increased nor decreased by introducing the notion that the state of the planetary system at one point in time is a consequence of previous conditions. The concept of cause is irrelevant

to the statement of the laws of the planetary system and is unnecessary for making predictions from the laws. The reader may here jump to the conclusion that the concept of cause has no role to play in the development of a science, but this conclusion is not justified. While the concept of cause may not enter into the statement of the laws that are the final products of the work of the scientist, it does nevertheless play an important role in his thinking and may help him in the discovery of laws. Even scientists in advanced fields of knowledge admit to *thinking* in terms of cause and effect. For example, biographical accounts of Einstein's early thinking illustrate this point vividly. A personal concept of the nature of the universe seems necessarily to involve some concept of the causation of events, and this way of thinking plays an integral part in the process of scientific discovery. A brief discussion of the origin of the concept of cause may make this point clear.

Personal experience leads one to believe that all events are the products or results of other events, which are referred to as their causes. The product of these causes is referred to as the effect. The origin of the belief in the existence of causal relationships is found in personal experience, in that each of us performs at least certain acts for the purpose of producing certain effects. In addition, we have the experience of other events producing certain effects on us. This has led ultimately to a concept of a universe in which every thing or event has a cause and in which there is thus a continuity and order among events. Certain major scientific concepts are deeply rooted in this concept of causation—for example the principle of the conservation of energy, which states that energy can neither be created nor destroyed. A similar principle deeply rooted in the conception of causation is the principle stated by Pasteur that only life gives rise to life.

While it is desirable to avoid the projections of one's own personal experiences onto the universe at large, there still seems to be merit in retaining a conservative conception of cause in thinking about natural phenomena. Indeed, it is almost impossible to conduct such thinking without the concept of causal relationships. This conception retains the idea that in order for a particular event to be produced, it is necessary for certain conditions to exist, and these necessary conditions are referred to collectively as a cause. Much of what

has been learned by scientists, such as the educational psychologist's concept of maturation, would be difficult to think about at this time without the concept of causation.

Some scientists prefer to state that they are seeking to establish systems of *functional* relationships rather than causal relationships. This must now be explained. Consider, for example, a simple and well-known law, such as Ohm's law, which can be stated in the following form:

$$\text{Potential difference} = \text{current} \times \text{resistance}.$$

This may be interpreted in popular language as follows: When a current passes through a wire, the drop in voltage along the wire is proportional to the product of the resistance of the conductor times the current. Now one cannot say that the potential difference is caused by the resistance to the current, for the causal relationships are complex. Nevertheless, the relationship expressed by Ohm's law represents interrelationships among phenomena, which, if fully described, could be represented by a set of causal relationships. These relationships could be described in terms of the electron theory of the structure of matter. A description of the system of causal relationships on which the law is based would be complicated, and much more elaborate than the law itself. For this reason, some scientist may prefer to say that Ohm's law represents a functional relationship among variables. Hence the reader will see that the term "functional relationship" refers to a situation in which is described a relationship that is not directly causal but is based on a complex system of interactions.

Most relationships in the behavioral sciences are not expressed in a form like that of Ohm's law or similar simple relationships. However, a few such simple relationships have been postulated—for example, that postulated by Hull (1943) in the form $R = HD$. When translated, this equation reads: Response evocation is equal to the product of habit strength and drive. Insofar as this simple equation expresses a true relationship, it does not represent a simple causal relationship but the result of a great complexity of relationships.

There is perhaps a certain safety in using the term "functional relationship" rather than the term "causal relationship." When we demonstrate that rewarding a child for performing certain acts in-

creases the probability that the child will perform those acts in the future, we may say that we have established a functional relationship. In doing this, we are avoiding making any statement that there is a direct causal relationship between the reward and the heightened tendency to perform the rewarded behavior. Nevertheless, the implication is that the relationships are more than just coincidental, but are a necessary part of the phenomena studied. We can be almost certain that the relationship between the reward and the changed response probability is an extremely complex one, not one that is well described in terms of a simple and straightforward causal relationship. Some philosophers would even go so far as to say that almost any relationship that appears on the surface to be a simple causal relationship is, in fact, a matter of great complexity. For this reason, throughout this book we propose to use the term "functional relationship" rather than the term "causal relationship." Of course, we will not quarrel with those who prefer to use the term "causal relationship" from time to time.

Knowledge Can Be Acquired at All Levels of Precision

There has been a tendency in psychology as it has developed within the American culture to emphasize the need for expressing theories in terms of variables that can be measured. For the most part this emphasis has been a healthy one, for it has led psychology away from the field of philosophy in which it had its beginnings and developed it as one of the biological and social sciences. In this matter, however, one can carry the emphasis on measurement too far. There are those who would prefer to quantify the trivial rather than to study the significant with qualitative methods that fall far short of the precision to which modern sciences aspire. It is perhaps worth reflecting on the fact that most of the generalizations of science began as qualitative statements and were later developed more fully in a quantitative form. For example, one hundred years elapsed between Newton's postulation of a universal gravitational constant and Cavendish's successful attempt to measure this important constant. The basic principles of thermodynamics were first stated in a qualitative form. Harvey's discovery of the circulation of the blood, together with the other qualitative discoveries of the great school of medicine at Padua, laid the foundation for what became ultimately the quanti-

tative science of physiology. The important discoveries that represent the very cornerstones of a quantitative science are almost invariably of a qualitative nature.

The writer is not advising the graduate student of education to plan qualitative studies for his master's thesis. Major qualitative contributions in research are made by the few, rather than by the many who make substantial but not brilliant contributions. It is almost essential that the graduate student build his research on the qualitative contributions and generalizations of others. He should appreciate the great importance of these qualitative generalizations and realize that the quantitative studies that follow build on the foundation which they have laid. Our present emphasis on quantification should not prevent us from perceiving its merits in their true light.

There is a place in every master's thesis and scientific paper where the researcher may attempt to formulate and present his own qualitative generalization. This is the section of the report that is usually titled "Implications," where the writer is generally free to give vent to his imagination.

Types of Laws

The scientist is able to make predictions when it has become possible to state a generalization, or law as it is commonly called. Laws may be either highly limited or broad in the range of events that they include. In this connection, the reader should note that two broad classes of laws have commonly been considered in the behavioral sciences, and these must now be considered.

The traditional goal of a science of behavior has been the discovery of laws that apply to all individuals; that is to say, laws that have wide applicability. Some psychologists have suggested that such laws may not be the only type of law that can be used in the development of a science of behavior. Indeed, some have even suggested that such laws may have only the most limited value. An alternative proposition, particularly highly favored by those who work in the clinical field, is that there are laws that pertain to the behavior of one individual but do not apply to the behavior of other individuals. Thus it is claimed that the sequences and orderlinesses of behavior manifested by one person may be entirely different from those manifested by another under similar circumstances. The order-

linesses of individual behavior that are unique to that individual are referred to as *nomathetic* laws, a term that distinguishes them from *ideographic* laws applying to all individuals. The clinician seeks to establish the laws of behavior of his patient so that he can predict how the patient will react to various possible modes of treatment and so that he can identify those aspects of the environment of the patient that should be changed in order to facilitate therapy. Nevertheless, the clinician also usually assumes that the unique aspects of a patient's behavior were generated through the operation of laws that apply to all individuals. He may assume, for example, that certain basic laws of learning may cause one individual to learn one set of habits and motives while the same laws may result in other habits and other motives in another person.

Allport was the first to raise this problem, in his book entitled *Personality* (1937). In this work, he suggested that different people were characterized by different traits, and that a major problem of the psychologist was to determine just what traits were operating in particular individuals. He recognized the difficulty of this problem and could suggest no satisfactory method of determining which traits were operating or how they could be measured.

One of his students, McKinnon (1938), demonstrated the reasonableness of this conception of personality in a study wherein he showed that, in handling a series of test situations, some individuals were consistently honest, others consistently dishonest and still others showed great variability of performance. The behavior of the last group could not be considered to reflect an underlying trait of honesty, but the behavior of those who were consistent could. However, the elaborateness of the technique needed to make a simple determination of the presence or absence of a single trait suggests that it is not feasible to determine by the McKinnon type of technique the presence or absence of a whole set of traits. Allport, in considering the problem, suggested that individuals might be grouped into types. Each type would consist of those whose laws of behavior were closely related. Thus it would be necessary to develop a technique for sorting individuals into types. But Allport pointed out that this solution was not a very satisfactory one, since it dealt only with approximations and necessarily introduced large errors in the prediction of behavior.

Interest in this whole problem was revived through the publication

of Stephenson's work on Q-methodology (1953). In this book a new approach is offered to this problem, and considerable material is also brought together from some of the older approaches. One of the techniques that he has invented is the structured behavior sample, which offers some hope of identifying which ones of a set of traits is useful in characterizing an individual. He also proposes to apply factor analysis to the problem of categorizing individuals in groups whose behavior can be understood in terms of particular patterns of traits.

These ideas of Stephenson should be considered at this time to be only explorations that may lead to the development of techniques that can solve practical problems. The techniques proposed are at present of use for research purposes and do not yet seem to be developed to the point where they can be used in the applied field.

A Program of Research Is a Long-Term Development

The discussion up to this point has been directed toward helping the graduate student of education develop a concept of the nature of research processes in education. The student may well wonder why large projects with which he is familiar often seem to end in relative failure. Such projects often have substantial backing; their failure is not due to lack of funds, and the probable reasons for this are worth examining at this point.

In recent years foundations have provided large sums of money for research on specific problems. An assumption on which such grants are based is that it is possible to formulate and plan at one time an extended program of research which can be undertaken more or less as planned. In the experience of the present writer, such projects in the field of education usually have been tragically ineffective up to the present time. The money has been spent, and almost nothing remains as evidence of accomplishment. The author can recall one such program with which he was associated as a graduate student and which dragged on for some years after he left it. It contributed nothing to substantive knowledge, and the staff slowly drifted away as funds ran out. Whatever it accomplished became relegated to the graveyard of forgotten and inconsequential events of educational history. The fact that memories of such projects carry with them all of the pain and anxiety of failure makes them all the more easily forgotten, so the lessons of failure are not learned.

The writer has pondered on this situation for long, and he believes that a substantial case can be made out for the existence of certain conditions that make it virtually impossible to develop at one time an extended program of research under the existing conditions. These conditions now are considered.

First, it has been stressed here that a unified program of research should be based on a common system of constructs and a unifying theory. Such conceptual systems cannot be developed in short order. They require an extended period of development. They may be derived in some initial and crude form from the literature describing previous research. The adaptation of the conceptual system so derived to the field under investigation, however, requires much experience with the field, and such experience comes only through close contact with the planning and execution of research. It is not enough merely to read about research, but the concentrated thought that should accompany the undertaking of research appears to be an essential ingredient of good theory-building. Thus actual participation in research is necessary for the adequate formulation of plans.

Second, a program of research develops out of a series of often loosely connected investigations, each of which is designed to explore one of several possible directions along which the program might develop. Such explorations are likely to be a necessary precursor of systematic and programmatic research. If approaches based on particular theories have been developed and appear to be profitable, then the way is clear for the development of a program of research.

Third, educational research has been taught to the graduate student of education as if it could be undertaken by the application of a simple formula. *This idea of research-by-formula has been the Nemesis of many major projects. It neglects the creative aspect of research, which is essential for genuine discovery and which is the heart of every significant research enterprise.* Research by formula becomes less and less adequate as the size of the program increases. The approach may be adequate for the small local investigation, but not for a large and continuing program.

Fourth, there is a matter that the writer believes has not received adequate recognition. It is maturity of judgment in deciding what is and what is not investigable. There certainly seem to be vast differences between individuals in their ability to identify problems that,

at the present time, can be investigated. Some researchers are extraordinarily "good guessers" in this respect, while others just do not seem able to perform this function. Undoubtedly, it is an ability that is highly dependent on experience. One is not likely to be a "good guesser" unless one has had wide experience in the undertaking and planning of research. Without such experience, the evaluation of prospective research projects is a virtual impossibility. The execution of successful independent research requires a lifetime of preparation. At the present time, education lacks a sizable group of high-level personnel who have made research a lifetime pursuit. Only through such individuals can large programs develop successfully.

Fifth, related to some of the previous points and yet in addition to them is the tendency for a researcher faced with the prospect of a large sum of money to overestimate what can be done with it. A sum such as $250,000 may seem to be immense to the poorly paid college professor, but the fact is that research is an expensive activity to undertake, and even the carefully planned use of a sum such as this may result in what seem to be relatively small corresponding gains in knowledge.

The large educational projects observed by the author that have ended without yielding useful knowledge have in all cases failed to develop according to the pattern that has been outlined. Rather have they been planned as grandiose ventures that have attempted to solve problems of central importance in the field of education through some comprehensive design laid down at the outset. Most of us have become much wiser in this respect over the last two decades—and perhaps much more modest in what we can aspire to discover in return for financial support for research.

Early Theory-Oriented Educational Research

The reader should not be left with the impression that educational research began by investigators collecting a vast amount of empirical information, followed by a period when theory has become an ever increasing influence—much as the data-gathering period in astronomy was followed by Kepler who organized the vast quantity of facts within a single and unifying theory. In education such has not been the case, for much early research was greatly influenced by what was then current behavioral theory even if it were not based upon it. Herbartian theory is easily seen in the work of Joseph M. Rice in his pioneer research on teaching. Rice, it will be remembered, had spent

a period of time in Germany, where he had become acquainted with new developments in pedagogical theory. This experience fired him to conduct his investigation on spelling and the relationship of achievement in spelling to certain teaching conditions, such as the amount of time devoted to drill. In addition, it must also be brought to the attention of the reader that when Edward L. Thorndike came to Teachers College, Columbia University, near the turn of the century, it marked an era of educational research dominated by the type of associationist theory that formed the basis of Thorndikean psychology. There was hardly an area of psychological educational research to which Thorndikean theory of learning was not applied. The learning of arithmetic, reading, and writing was scrutinized in terms of this theory, and thereby a great quantity of knowledge was acquired about these processes.

Research of a so-called fact-finding nature, which is typical of much that is done today, represents a phase that finds its roots in the 1920's, with the growth of research departments within the big-city school systems and within the educational departments of the various states. Such research was intended to solve the various problems that were constantly arising and that needed to be solved in order to be able to map out effective educational policies. It is one of the misfortunes of our times that the policies of such research departments were often set by men in high position who had little understanding of what research workers could or could not accomplish. Such individuals were largely unaware of the research worker's purpose as it is stated in this volume, but rather did they regard him as an inventor of gadgets and knickknacks and an unearther of odd and peculiar facts. With this type of "research" policy in mind, research offices were staffed with workers interested in devoting their lives to the collection, tabulation, and interpretation of facts. Under this influence, educational research became a massive fact-finding enterprise. It is hardly surprising that the resulting large-scale projects have failed to produce generalizations that have added to our knowledge of the laws of behavior in educational situations. Indeed, the provincialism that underlies this domain of research is designed to produce local answers to local questions.

Summary

1. Most educational theories do not form an adequate basis for research. They tend to be vague and do not specify what conditions produce

what. Many fail to identify the variables that may be involved and hence give no cues as to what should be measured.

2. Theories developed for the purpose of providing a basis for a program of educational research probably should refer to limited phenomena. This means that they should be fairly simple. Complicated theories may have their place when the science of behavior in educational situations is at a more advanced level than it is today and when the variables can be measured or evaluated in a meaningful way.

3. Illustrations are given of attempts to state theories related to limited aspects of education. Such attempts are primarily ways of organizing one's knowledge so that what is known can be clearly seen and so that gaps and deficiencies can be noted.

4. The present writer believes that the educational research worker should attempt to establish functional relationships among events and not merely statistical relationships.

5. The research worker in education should attempt to add to the body of knowledge that already exists in the field. He should not seek to contribute isolated items of information that may be interesting in themselves but that do not contribute to an organized body of knowledge.

6. The present need for educational research to be based on explicitly stated theories represents, to some extent, a need to return to an earlier period when theory-oriented research was the rule. The impressive contributions of such persons as Thorndike early in the century are convincing evidence of what theory-oriented research may be expected to achieve.

7. Knowledge can be acquired at all levels of precision. In the early stages of inquiry, knowledge about phenomena is likely to be very inaccurate and general, but it may be of immense importance for later developments.

Some Problems for the Student

1. Study John Dewey's theory of problem-solving in his book *How We Think* (1910). On what kind of information is this theory based? What kind of deductions can be made from the theory to test its validity? Why is the theory not very useful as a basis for research?

2. What kinds of deductions that could be used to test its validity can be made from the Montessori theory?

3. Examine a textbook that describes procedures for teaching children to spell. Draw up a set of postulates that describe the general theory on which the procedures are based. A similar exercise can be performed in other curriculum areas.

The Content of Educational Research | 3

The Broad Areas of Educational Inquiry

Educational research as it is known today is a relatively new branch of knowledge. Little more than half a century has elapsed since Joseph Mayer Rice planned his researches on the teaching of spelling and other skills, and hoped through his studies to bring reform to education. While research has not yet become the tool for educational reform which Rice conceived it to be, it has some substantial changes to its credit. The reforms brought about by research have not matched expectations because the best part of a half-century has been taken to find out a little about how to do research in education, what can and what cannot be accomplished with the crude tools at the disposal of the researcher. The relatively little that has been accomplished has been the product of great effort. Nevertheless, educational research is now beginning to pass beyond the adolescent stage of ambitious but unrealistic dreams and to conceive of its role in more mature terms.

Although accomplishments have been much less than early devel-

opers hoped, research has had substantial effects on education. In the lower elementary grades, we find today readers that have been carefully designed so that each new word is presented a sufficient number of times to allow the child the opportunity for learning it adequately. Such readers have been developed on the basis of learning studies. The measurement of the pupil's reaction to reading is determined by means of tests that are the products of extensive research. So, through the grades, the influence of educational research is evident in the techniques that are used. However, research still provides only meager advice concerning the way in which the teacher should manipulate the classroom situation in order to maximize learning or to produce specific results.

Educational research, as it is conceived here, represents an activity directed toward the development of an organized body of scientific knowledge about the events with which educators are concerned. Of central importance are the behavior patterns of pupils, and particularly those to be learned through the educational process. A scientific body of knowledge about education should enable the educator to determine just what teaching and other learning conditions to provide in order to produce desired aspects of learned behavior among young people who attend school. Presumably, learning conditions will also have to be suited to the aptitudes and other characteristics of the learner. Where the researcher can most advantageously begin to develop such an organized body of knowledge about educational events is still a matter of conjecture. He may decide to begin by studying pupil behavior itself, or by studying conditions that affect the pupil only indirectly, such as economic conditions and school finance. Wherever he does begin, however, the assumption is made that the phenomena studied affect in the ultimate analysis the pupils in the schools. Thus our study of the methodology of educational research begins by taking a brief overview of the kinds of problems that are dealt with by those engaged in this field.

Educational Research Related to Development

Studies of the sequence of events in the developmental process have long formed a topic of central importance in the educational field. The extensiveness of research in this area is seen in the fact that the *Review of Educational Research* devotes an entire issue to

the area once every three years. The 1955 issue covered over 850 references to research. To state the over-all purposes of such research is not easy because they are so varied. Nevertheless, studies can be conveniently classified into two major areas. First, there are those concerned with development over rather long periods of time, such as a year or several years. Second, there are those concerned with development over shorter periods of time. The latter studies usually are undertaken in a school setting and are primarily concerned with learning as it occurs in school. Let us now obtain a brief overview of the first-mentioned group of studies.

Long-term developmental studies find their roots in biology, and for this reason they have been dominated to a considerable extent by the maturationist point of view. The work of Arnold Gesell on the development of behavior in the infant and young child is typical and seems to have been directed toward the discovery of laws of development. Much the same is true of the classic work of Jean Piaget and his associates in their studies of the development of number concepts and space concepts in the child. Implicit in much of this work is the notion that the sequence of events in the developmental process is firmly established by the inherent nature of the organism, and that the laws of human development can be described in much the same way as the sequence in a complicated chemical reaction. A part of the apparent stability of events in such studies may be due to the fact that the children studied formed a relatively homogeneous population and had similar learning experiences. Experiments designed to assess the consequences of a systematic variation of environmental conditions are difficult to accomplish in the case of the human child. While Gesell is often thought to represent the maturationist point of view, it is much to his credit that he was one of the first to initiate experimental studies in which the environment of the developing child was systematically varied. Just how much variation in the developmental pattern can be produced by environmental changes still needs extensive further exploration.

Much of the work on development that has just been described seems to derive its impetus from the Rousseau type of educational philosophy, which emphasizes development from within, where the purpose of the educator is to provide an environment that does not inhibit this process. Such an outlook has become more and more out

of keeping with modern thought. The trend has moved toward emphasizing the effect of particular environmental conditions on learning.

Another type of long-term developmental study is directed toward an entirely different purpose, namely establishing the changes that occur during maturity and the later years of life. To a great extent this work explores the educability of adults and the problems that they face in a rapidly changing economy wherein many find their jobs becoming obsolete and are forced by this fact to find new forms of employment.

Still another type of long-term developmental study has appeared in recent years, which reflects some of the statistical sophistication that psychology has begun to acquire. In this type of study an attempt is made to determine the change in the structure of the psychological factors that appear in tests when they are administered to groups at various age levels. A large number of studies of this kind have appeared in recent years, and they have generally tended to the conclusion that an increasing number of distinct abilities appear as age increases through childhood. It would be of immense interest to the educator to know whether the educational program can affect the emergence of these factors.

Finally, a type of developmental study that has recently found prominence in the literature attempts to determine the pattern of physical development for the purpose of providing a basis for the design of school equipment such as tables, chairs, and desks. This might be termed the human-engineering approach to educational problems. The application of this approach is difficult in the educational field when it is realized that, despite wide variation in the physical characteristics of any age group, all pupils have to be accommodated in the school. In contrast, the human engineer may often design his equipment so that it can be operated by only 60 per cent of potential users.

Another problem that is raised by studies of growth in relation to equipment design is that there are not only large differences between pupils of the same age, but that these are related to differences in sex, in locality, and in other factors. Equipment therefore must be chosen for particular groups of pupils, not for particular age groups.

The studies of development considered up to this point have been dominated by the maturationist point of view. Long-term studies have

been very much of this character—but, of course, with many notable exceptions. In recent years there has been an increasing realization that developmental studies should explore the influence of specified environmental conditions on human learning and behavior, and in particular the effect of aspects of the school environment. With this realization, there has been an increasing tendency to explore the relationship of development observed among children to variations in environmental conditions. Practical considerations have tended to make this second group of studies short term rather than long term. The classroom and events therein define a traditional area of research for educational psychologists. Obviously, a search for classroom determinants of pupil behavior is of immediate importance to the teacher and school administrator. Such research is directed toward the discovery of laws that relate teacher behavior and other events in the classroom to those aspects of pupil behavior that the educational process is designed to develop and produce. Studies of the relationship of characteristics of teacher behavior to changes in pupil behavior would fall in this category, and so too would studies of the relationship of textbook characteristics to the rate at which pupils acquire proficiency in reading or in some other skill. Fortunately, there exists in the field of learning a substantial body of theory that can form a basis and framework for such studies, and hence knowledge in this area can be developed in a fairly orderly and systematic way.

Since the behavior that the educational process seeks to generate in pupils is a product not only of events occurring in the classroom but also of the personal characteristics of the students, studies of the relationship of individual differences to the facility with which various skills or other characteristics are acquired are commonly undertaken. Social phenomena also come within the purview of research in this area, since socialized learning is commonly introduced in the classroom in a variety of different forms.

The term *learning* covers a multitude of phenomena. In the past, the emphasis of educational psychologists has been on studies of the learning of skills such as reading, spelling, and computation. The reason no doubt was that the available framework of theory was helpful in the formulation of studies of learning of that type. The Eight Year Study made a significant departure from this tradition by

the initiation of influential studies of the acquisition of certain aspects of thinking skill. Since that time there has been an increasing emphasis on the study of the higher mental processes and the conditions related to the acquisition of reasoning and other high-level skills. Theoretical developments have also been made that have aided greatly the formulation of programs of research in this important area of education.

Sometimes the educational psychologist explores conditions, more remote from the immediate educational situation, that are hypothesized to influence the acquisition of new behavior in the school situation. For example, differences in home background result in differences in the experiences that pupils bring to the classroom. This in turn may facilitate or interfere with changes in behavior that the school is attempting to produce.

While the traditional emphasis in education has been on intellectual development, and there are many still who wish to retain that emphasis—some even advocating that it should be the only emphasis—the trend has been to consider that the school has responsibility for the development of the personality of the pupil. Early studies in this domain were concerned mainly with the role of the school in the development of rather superficial personality characteristics as represented by attitudes and interests. More recently attempts have been made to study characteristics of deeper significance. For example, studies of the origins of aggression have shown considerable promise. Other studies of achievement motivation appear to be concerned with a characteristic of the greatest importance in the educational process and one that can now be successfully measured.

Research on personality development can be undertaken within a number of different theoretical frameworks. Traditionally, much research has been focused on problems of developing desirable personality traits in pupils, particularly those personality traits supposedly related to adjustment. Another approach is that personality is as much a reflection of the situation in which the pupil is placed as it is a consequence of an inner and hidden pattern of traits. That a position of this kind has some validity is seen in the fact that the teacher who may give the appearance of being a severe and cold person in the classroom may, in her social contacts, be the life of the party, with a warm and genial personality. Insofar as this is the case,

school practices should be such that they elicit only those aspects of pupil personality that it is desired to cultivate. In this connection, studies have been undertaken that attempt to relate the social behavior of the teacher to the social behavior of pupils. For example, studies have demonstrated that aggressive behavior on the part of the teacher results in aggressive behavior on the part of the pupil. Much can be done within this theoretical framework, for the way seems to be open to the study of the relationship of manifest teacher behavior and manifest pupil behavior in the classroom. Such studies should provide important information for teacher education. Perhaps a sobering note may be introduced by the thought that the behavior of the teacher is possibly at least in part a product of pupil behavior.

Many other theoretical frameworks have formed the basis of other studies. There is a wealth of ideas on which the educational research worker may draw. No longer does research in this area have to be a mere accumulation of facts, for it can systematically contribute to a growing body of knowledge.

Research on problems of classroom learning varies in the extent to which it may be applied directly to the teaching and the guidance of pupils. In the hope that the results may be directly applicable, many studies are conducted in the classroom milieu, with the minimum interference with class routine. Such studies often present serious difficulties when the results are to be interpreted, since uncontrolled events are numerous and often cannot be identified even though they may influence the outcome of the study. The difficulties of studying learning phenomena in the classroom are such that many research workers have felt that studies of learning can be better undertaken in the laboratory. Here rather small segments of behavior are studied, but under very carefully controlled conditions. A great amount of what is known about the learning process of the child was derived from such procedures. The work of Edward L. Thorndike, which still exerts an influence on educational practice, was undertaken largely in a laboratory.

The writer does not think there is merit in arguing whether the laboratory or the classroom is the better place for research. Each has its own merits and advantages. The vigorous development of educational research requires that studies be undertaken in both of these settings. Current research shows that both avenues are being explored.

Curriculum Research

The term *curriculum research* covers a multitude of very diverse activities. This is partly because the concept denoted by the word "curriculum" has had an evolving and expanding meaning and curriculum research has shown a corresponding evolution and expansion. While a century ago the concept of a curriculum was that of a body of subject matter to which the pupil was exposed, the present-day concept is largely a different one although the old-time meaning has not entirely vanished.

In Carter V. Good's *Dictionary of Education*, published in 1945, the following three distinct meanings of the word "curriculum" are noted:

1. A systematic group of courses or sequence of subjects required for graduation or certification in a major field of study, for example, social studies, physical education curriculum.
2. A general over-all plan of the content or specific materials of instruction that the school should offer the student by way of qualifying him for graduation or certification or for entrance into a professional or a vocational field.
3. A body of prescribed educative experiences under school supervision, designed to provide an individual with the best possible training and experience to fit him for the society of which he is a part and to qualify him for a trade or profession.

While the first and the last of these definitions involve the notion of a body of content, the second definition introduces the idea that an *over-all plan* is also an essential feature of a curriculum.

Further light is thrown on the development of the concept of a curriculum in the *Review of Educational Research*, Volume XXVI, Number 3, 1956, prepared on the occasion of the twenty-fifth anniversary of the founding of the *Review*. This particular issue traces the development of curriculum research and points up some of the changes that have taken place in the educators' concept of a curriculum. It is implied in this review that the tendency in the past was to think of the curriculum as consisting of all the *experiences* that a pupil had during schooling. The emphasis according to this outlook was on the experience aspect, and hence the curriculum was considered important insofar as it represented an element in the pupil's conscious experience. This was, in fact, a very narrow conception of

the curriculum, for surely there are important factors in the pupil's environment that have powerful influences on his behavior but of which he is never aware. Any useful conception of the environmental influences that play a role in the development of the child must not be limited to those to which he consciously responds. For this reason, the emerging concept of the curriculum held by research workers and others is that it consists of all the planned conditions and events to which the pupil is exposed for the purpose of promoting learning, plus the framework of theory that gives these conditions and events a certain coherence. Recognition that a framework of theory is needed to give meaning to the happenings in the school is an important step forward. Virgil Herrick and other modern writers on this subject stress that a curriculum can have no real meaning unless it is part of a theory of education.

The introduction of the idea that a theory is a central and essential aspect of a curriculum has been an important step in the development of curriculum research. If one looks back over the research accomplished in the area, one is impressed by the fact that the chief weakness presented is a lack of any theoretical foundation. There was a tendency to study the effect of one set of materials in comparison with the effect of another set of materials. Analyses were made of the content of textbooks. One procedure was compared with another. Sometimes a clearly stated curriculum theory was presented, but most often it was absent. The current emphasis on the need for stating clearly curriculum theory should do much to remedy this defect in curriculum research. Indeed, the effect of this trend in thinking is already noticeable.

Techniques for the study of curriculum problems exist to some extent, and these have to be wedded to a body of theory. For example, there are already many methods for measuring many important characteristics of textbooks. Readability is one of the best explored of these characteristics, and while much needs to be done to improve readability formulae, those at present available are still valuable research tools. There is a need for techniques that will measure the complexity of written materials. An excellent beginning has been made in the development of techniques of content analysis, and enough has been accomplished to inspire research workers to apply these techniques to the study of curriculum problems. Much less developed are the techniques for measuring the important character-

istics of the other curricular materials to which the pupil is exposed, such as excursions, visual aids, and workbooks. In recent years considerable interest has been shown in studying the value of radio and television for educational purposes. Extensive research in this area has resulted in the development of many techniques that can be used as the tools for further studies.

A particularly difficult area of curriculum research is that involving the actions of the teacher and the relationship of his personality traits to his classroom behavior. This is an area largely lacking in theoretical constructs, although considerable work, designed more to explore than to test hypotheses, has been initiated on an empirical basis. In the older studies attempts were made to determine the relationship between the measured or rated personality traits of the teacher and "teaching effectiveness" as measured by ratings. Since it is becoming more and more apparent that there are difficulties in measuring teacher effectiveness, more recent efforts have been devoted to discovering relationships between the personality of teachers as measured in various ways and characteristics of teacher behavior as they appear in the classroom.

Numerous other curriculum problems have formed the basis of various inquiries. Studies of pupil needs in relation to curriculum design are legion, and so too are studies designed to discover the extent to which various skills can be learned at particular levels of maturation. Historical and sociological studies also constitute an important part of curriculum research, and so do studies of the social and political forces that influence the curriculum.

Research Related to Sociological and Economic Conditions Affecting Education

Many of the conditions external to the classroom that affect education are investigated by educational research workers interested in studying the sociological conditions that ultimately influence the educational process. Such a researcher might study the political pressures that influence the educational philosophy of a school system in one direction or another. He might also view the financial aspects of the system and the sociological conditions that result in increased or decreased support for education. He might seek to answer the question, "What philosophy of education is implied in the curriculum

of the school?" or, "What kinds of individuals constitute school boards?" or "What barriers exist to the raising of funds in communities of certain specified types?" The answers to such questions supply information that can be applied only indirectly to improvement in the effectiveness of learning in schools. Many such studies attempt to answer questions of local significance only, and the results cannot be applied to other communities. Such studies result in the production of what may be termed low-level laws.

While the central topic of educational research is the development of a science of behavior in educational situations, peripheral studies pertaining to sociological and anthropological problems also are of considerable importance. Studies of the latter type and those pertaining to behavior in the classroom can be considered to be on different levels of description. Just as classroom behavior may be described at various levels, from those that involve the movement of the constituent elements or muscle twitches of the body to descriptions of the over-all behavior of the individual, so too can one conduct studies either at the level of individual behavior or at a level where the unit studied is a group. Sociological and economic studies of educational problems are concerned more with group phenomena than with individual phenomena. To some extent, historical studies are of this type. Studies of school finance are of this character, and at the same time they present many of the difficulties inherent in sociological approaches to educational problems.

Sociologists, economists, and anthropologists use quantitative methods in their respective areas in a way rather different from that of the psychologist. Only rarely can the former group of scientists set up experiments that provide simplifications of the conditions that actually exist. Without experimentation, the scientist is faced with a vast multiplicity of phenomena, which do not always lend themselves to arrangement in a metric scale. Nevertheless, present-day sociologists and economists, and to a lesser degree anthropologists, are developing quantitative approaches to the study of human behavior that could be adapted to the ends of educational research.

Even if one turns to a relatively well-structured field such as school finance, where one would expect to find that phenomena were well quantified since much of it appears to deal with quantities, it is quite clear that researchers have struggled long to develop useful variables.

The reader may well be referred at this point to a work by Mort and Reusser (1951), which illustrates the problems of obtaining useful measures that can be used by research workers. A particularly informative example is presented by their chapter on educational need, which discusses attempts to develop measures of the educational requirements of communities. These measures must be such that they are comparable in some way from community to community. At first sight one might be tempted to take the pupil as the unit of need, but it soon becomes evident that this is not a satisfactory procedure. If the pupils are widely scattered, schools tend to be small and the cost unit greater than when schools are large. Scattered pupils also increase the cost of transportation to the school, and then the pupil unit has to be adjusted for such factors if it is to represent a unit of educational service to be provided. These complexities have resulted in the development of an extensive technical literature, which has attempted to derive a useful measure of educational need that could be used in subsequent research.

Studies of the economics of education often provide useful data for the solution of immediate problems, but rarely do they provide generalizations useful for the solution of a great range of problems. The generalizations that can be derived are qualitative and usually lack any property that might be described as precision. Paul Mort, who has thought through this matter at great length and perhaps with greater depth of understanding than any person now living, points out that these generalizations are not scientific laws but rather are they categories under which may be classified a "multitude of rule-of-thumb canons," which are widely accepted as the major lessons that have been learned through studies in this area. One of these is the "equalization" principle, which states that financial disbursements are such that poorer districts are helped more than wealthy districts. Another is the reward-for-effort principle, which is that funds are disbursed in ways that encourage other bodies also to release funds for education. An example of the operation of the latter principle is the case where states provide educational funds for local communities on condition that the local communities expand their educational expenditures by a certain amount.

In our opinion, Professor Mort is perhaps too modest in the description of these and other principles. They represent low-order

generalizations that are essentially scientific in nature. When one considers the limited degree to which this area of educational research has developed quantitative methods, it is gratifying to note the order that such principles have apparently introduced.

An area of educational research dependent upon sociology is demography, that is to say, the study of population changes. Its importance to education is substantial from many points of view. First, if the supply of teachers is to be related to need, then it is necessary to predict several years in advance the number of classes of a given size that are to be staffed. These long-term predictions are required for adequate educational planning because it may take four or five years to train a teacher and perhaps another two years to recruit him in the first place. In addition, some experience on the job under qualified supervisors also seems to be most desirable. Thus, if there is to be an increased number of teachers available in 1980, it will be necessary to start an active recruiting campaign possibly as early as 1970.

In demographic studies, there are no great difficulties in estimating pupil enrollment five or six years ahead since the future pupils already have been born at the time when the estimate is made. Death rates and immigration and emigration rates can be estimated for this group with considerable accuracy. On the other hand, real difficulties are encountered if estimates are to be made ten or more years in advance, for the birth rate may show sudden changes and these may result from a great complexity of causes. Hardly more than a beginning has been made in establishing scientific laws of population change. Just as the prediction of pupil enrollment is an area fraught with difficulties, so too does the problem of estimating future teacher supply present numerous unsolved problems that may challenge the student of education.

It must be pointed out that these demographic studies have only the most limited value, largely because school authorities cannot recruit enough teachers at the present time and little control can be exercised over the flow into and out of the profession. The latter movements are controlled far more by prevailing economic conditions than by any action that those in charge of teacher education make take. There is, perhaps, only one aspect of this branch of the sociology of teaching that has any practical consequences. Through the information derived, it may be possible to persuade students in training to enter

those phases of teaching in which there is the greatest need for new teachers and to avoid the few areas where there may be an oversupply.

Educational Engineering Research

It is well known that in industry there has developed a branch of knowledge, referred to as "human engineering," which serves the purpose of adapting machines to the human operator. This problem has already been touched on in the discussion of developmental studies. If a body of knowledge exists about the requirements, abilities, and physiological properties of the human operator, machines can be designed that are well within his capacity. The design of equipment to conform to anatomical, physiological, and psychological requirements is an important aspect of industrial design and is now backed by a substantial body of knowledge. In education, there are certain types of related research that must be considered at this point.

First, there is the problem of designing the regular educational equipment such as chairs, tables, lockers, toilets, and the other minimum essential school furnishings. A study designed to provide a sound basis for the development of such equipment is one by Tuddenham and Snyder (1954). Such studies involve difficulties, which are considered in the section devoted to developmental studies.

Second, there is the problem of the design of educational equipment that is more directly related to the learning process itself. The field of education has been notoriously deficient in the development of mechanical devices that facilitate learning, but, as Sidney Pressey has pointed out on many occasions during the last twenty years, the industrial revolution will ultimately come to education and the classroom will then be equipped with mechanisms for the facilitation of learning.

It is too early to say what course the ultimate mechanization of the classroom may take, but there is plenty of room for invention at this time. As a matter of fact, mechanical devices have already shown a considerable development over the years and have become more and more suited to the achievement of acceptable educational objectives. Sidney Pressey pioneered this field, and over fifteen years ago he developed for the United States Navy a series of mechanical devices that could be used as trainers for the acquisition of informa-

tion and also as means of measuring the extent to which such information had been acquired. In these early machines the student looked through a window in the device and read a question. He then selected from a series of possible answers presented the one that he thought was correct and pressed the key corresponding to the number of that answer. If he were right, the next question appeared. If he were wrong, he could try another answer. The machine could be set to score the number of incorrect attempts or the number of correct answers.

Such machines are clearly of value only for learning certain types of factual information. Since each learning or testing series can be repeated as many times as the student may wish, he can plot a curve of his performance. At each attempt he may try to beat his previous showing. The device gives both immediate knowledge of results when each response is made and also a total score giving over-all performance.

Nobody would claim that such a machine as that developed by Pressey has utility except in a few situations in which strictly factual information is to be acquired. It is not designed to play any major role in any particular curriculum, but for its limited function it has merit. B.F. Skinner, who was familiar with the earlier work of Pressey, set himself the more ambitious goal of developing machines that would be closely tied to the curriculum of the lower elementary school grades. Skinner believes that it is most important that the pupil generate his own response rather than recognize as correct one of several presented. A second principle built into Skinner's machines is that the problems presented should not be just an unrelated series, as they were in earlier machines, but rather should form a sequence that builds learning in an organized and planned manner. The machine should schedule learning in some way that is efficient, and the schedule should be adapted and changed to take into account the abilities and maturational level of the pupil. Apart from the ingenuity and vision of the proposal, Skinner's work in this area represents one of the most concerted and systematic attempts to apply learning theory to problems of classroom teaching. One of Skinner's machines is shown in Figures I and II.

The reader should not jump to the conclusion that mechanical devices are of value only in the learning of simple skills, for such is

not the case at all. They also offer considerable promise as devices for providing both instruction and experience in problem-solving, decision-making, thinking, and related activities. Since 1955 a num-

Figure 1. Preparing to use a learning machine. A student inserts a disc bearing questions into a teaching machine developed by Professor B.F. Skinner of Harvard University. The machine is intended for use in high school and college classrooms and will make use of facts about learning that have been discovered in a twenty-five-year laboratory and classroom research program. Photo by courtesy of Professor Skinner.

ber of patents have been issued for machines designed specifically as training devices for tasks involving complicated thinking. The better-known of these devices have been designed for providing experience in trouble-shooting in complicated equipment. Others can be

adapted to provide experiences in problem-solving in social studies, in scientific areas, and in mathematics. This is an area for pioneer educational research.

Figure 11. Learning by means of a machine. A student uses new teaching machine designed by Professor B.F. Skinner of Harvard University. Student reads question in left window, writes answer in right window. If answer is correct, then machine may present next question; if incorrect, new question does not appear. Photo by courtesy of Professor Skinner.

Institutional Research

It is of interest to consider at this time the function of what has been termed institutional educational research and its value in developing a scientific body of knowledge. Research of this type has a

history that goes back more than a quarter of a century. The discussion here refers to the small educational research units that have been established in many large universities for the purpose of solving their own educational problems. A pioneer in this endeavor has been the University of Minnesota, which since the early 1930's has attempted to gain knowledge about its own educational problems through a small department developed for this purpose. Such departments have usually been staffed by both professional research workers and graduate student assistants. Typical topics for the reports that have been produced are concerned with such matters as the knowledge of the student at various stages in his career, his problems after he leaves college and the relation of these problems to the education received, the development of methods for establishing an equitable salary scale, the utilization of space within the school plant, the development of procedures for promoting faculty, and the reasons for failure to graduate. This sampling of topics is given because it reflects the element of administrative expediency that determines whether one study or another is undertaken by the unit. It also illustrates the close relationship that must exist between the institutional research unit and the administration of the university. To facilitate their cooperation, it is of advantage for the unit to be attached to the general administration of the university rather than to one of the teaching units.

In appraising what can be accomplished by such units, let us start by considering their merits. First, they represent a rational approach, not an arbitrary one, to the solving of major and pressing educational problems. They emphasize in a most healthy way the need both for the collection of facts and for reflection concerning these facts in the solution of educational problems, in contrast to the method of opinion and argument that is the common way of solving such problems. To many a member of a liberal arts faculty, the methods of a research unit represent entirely novel approaches to the solution of educational problems. The new approach offered by the research unit provides stimulation for thought and sometimes even makes for a realization on the part of the faculty that traditional methods of solving educational problems should be discarded. Once the latter has been accomplished, the faculty should attempt to define its problems in terms that are investigable and much of the characteristic vagueness of liberal

arts thinking about educational problems should be dispelled. The educational research unit should provide an adventure in thinking for the faculty.

Of course, not all such units have served this worthy purpose. The head of the unit must have not only the research skills necessary for executing the studies that are initiated but also a certain political acumen. Judgments have to be made concerning what ideas will be and will not be acceptable to the faculty. The researcher who attempts to cast doubts on the most cherished educational ideas of the faculty is likely to be soon seeking another position, but there are some professional researchers who have been extremely successful in providing a stimulus to the thinking of a faculty.

Such is the positive side of the picture, which presents a strong case for the development of such research units. In terms of the function they fulfill, institutional research units have justified their purpose. But what is their role in contributing to a body of knowledge to which other scientists contribute? The answer the writer would give to this is that only rarely have such studies made a contribution to scientific knowledge. The reasons are several, and they do not reflect in any way on the competence of the work that has been done by institutional research units. The writer believes that genuine scientific work can be undertaken while working on practical problems. However, such scientific contributions are likely to be made where the research is programmatic and where a particular type of problem is persistently attacked in an integrated series of studies. In institutional research, administrative pressures are likely to be such that as soon as one study is completed, another study in an entirely different area must be started. By this process institutional research solves numerous separate and isolated problems, some of which probe a subject to the point where a real contribution to knowledge can be made. To the administration the results are often immensely useful, but to science there is unlikely to be any permanent contribution.

The writer's opinion is that units undertaking institutional research should attempt both the answering of problems posed by the administration and the building of a body of knowledge about some problem of central and scientific importance. For this to be done, it is necessary to free some of the labor available for concentrated programmatic research and to protect this aspect of the work from pressures to han-

dle local problems. If this is done, it is believed that slowly there can be built up a substantial program of research related to an important educational problem, and it is likely that this body of knowledge may be of greater use to the institution than are the answers provided to day-to-day questions. Once a program has been initiated, it should be possible to implement the resources of the research unit with funds obtained from foundations that sponsor basic research. That "basic" research on educational problems can be undertaken along with "applied" problems is a position that institutional research units should take if they are to grow and flourish.

Areas of Educational Inquiry Beyond the Scope of This Book

Educational research conceived as a means of developing an organized body of scientific knowledge about those aspects of behavior that concern the educator clearly does not include all areas of educational inquiry. The philosopher who studies problems related to values in the field of education is not attempting to build a science of behavior in educational situations, and his approach differs from that of most scientists. It is quite beyond the scope of this volume to offer the student help with the development of inquiry in philosophy. The important key problems in education that the philosopher attempts to solve are clearly not those with which the scientist has had any success in solving. The limitations of the methodologies of the scientist should be recognized by the reader.

This book also does not touch on the methodology of historical inquiry, although a substantial portion of what is commonly described as educational research falls into this category. By *historical research* is meant the use of records of previous events for the purpose of arriving at generalizations that may be used for the solving of current problems. It is not only the educational historian who uses this approach, for it is common to find that the curriculum specialist often employs this same technique. For example, in curriculum research a student may seek to design a more effective plan for using the school excursion as a teaching device. He may do this by gathering together from the literature all the information he can find about the use of school excursions, and all the evaluative statements that have been made about them. On the basis of these historical data, he may develop a plan that incorporates what are believed to be all of the

effective uses of the excursion. The weakness of this method as pursued by the amateur is not always recognized, and therefore it should be pointed out here. The professional historian is, of course, fully cognizant of these problems.

First, as the writer sees the method, it is one that requires the researcher to discern certain sequences of previous events that appear to repeat themselves. Second, the method assumes that from such reputed sequences of events the researcher is able to derive generalizations that will permit him to make predictions about the future. Now this procedure is theoretically quite satisfactory, but difficulties arise because many compromises must be made in carryng it out. One compromise results from the fact that events occurring outside of the laboratory rarely are repeated under similar circumstances. The user of the historical method must make judgments as to whether a sequence of events has or has not been repeated. Thus the data available to the researcher who uses this method are far from satisfactory, and their interpretation is inevitably contaminated with the individual's judgment. For this reason the generalizations derived from such data must always be considered tentative.

The commonest misuse of the historical method is the employment of single instances for drawing generalizations. For example, suppose our student who was studying the use of the excursion by the historical method encountered in an educational magazine a description of an excursion in which the same place was visited more than once, and the writer of the article described the venture as a highly successful one. There would be no justification for the student to make the generalization from this single instance that repeated visitations are necessarily preferable to one-time visits. It might have happened in the case described that the nature of the excursion made it desirable to repeat the trip. The error of generalizing from a single reported event is perfectly clear in the example discussed, and yet it is an error that the historical researcher can make all too easily unless he is constantly on his guard against it.

On the positive side, historical studies may form the basis for planning scientific research of the type that is discussed in this book. They may form the basis of a theory, and deductions from that theory may provide hypotheses for subsequent inquiry. In this sense, the historical method parallels closely the procedures discussed here.

However, it is in the testing of hypotheses that the historian has difficulty in finding suitable data.

The scientific student of educational phenomena should have respect for the contribution that the historian can make to his work and should not brush it aside lightly. Only the most unwise educational research worker would proceed with a study without first determining what history had to say about the problem. Many studies in the curriculum area would never have been started if proper attention had been paid to the knowledge that history provides about the problems they attack.

The reader should also be aware that historians have their own special techniques for finding, selecting, and utilizing the facts of history. The unsophisticated should not venture forth and undertake historical studies without expert guidance from persons who have had long experience with such materials. While previous paragraphs have intimated that the historical method cannot do most of the things that this book aims to help the educational research worker to do, historical studies require great expertness to be undertaken successfully. The historian, like the scientist, is a specialized expert.

Levels of Research

Educational research, like all other types of research, may be undertaken at various levels of complexity, which range from the development of a simple technique or the reproduction of an experiment previously undertaken elsewhere to high-level discovery that is said to be the product of a creative mind. Some advice can be given to help the student at the simplest levels, but almost no help in words can be provided for the person who wishes to make discoveries. The research areas that have been discussed can be attacked at all levels.

In the hands of the highly qualified researcher, the development of techniques and the making of discoveries go hand in hand. He may supervise a program that involves all levels of research at once. The existence of a flexible technique often permits the investigation of innumerable different problems. For example, the development of objective achievement tests or sociometric techniques opened up ways of investigating all types of educational problems. The development of attitude scales permitted the study of the effect of a great variety

of educational experiences on attitudes. The development of word counts and measures of reading difficulty opened up new avenues to the study of the teaching of reading.

Often the development of techniques requires originality and invention, particularly where new areas are to be explored. For example, the Eight Year Study evaluation group directed its energies mainly to a large-scale effort to develop methods for measuring certain major outcomes of general education. Research on techniques is often important before much of significance can be discovered.

In recent years there has been a tendency to devote creative effort at a high level to the matter of developing techniques in the behavioral sciences. Perhaps this is a product of many discouraging results in studying problems of behavior with available devices, and of a feeling that until more adequate tools are developed, only limited discoveries can be made. In a sense, such work is peripheral rather than central to the work of the scientist. Technique development is not an activity at all peculiar to the behavioral sciences. The electronic specialist may become involved in the most complex mathematics in order to devise a piece of apparatus that the physicist may need for a particular experiment.

It is perhaps important to point out here that in the history of science, technique development has often been confused with the core scientific activity of discovering laws. Excellent illustrations of this are the procedures known as factor analysis and related techniques. Those who developed these techniques seemed to feel that the ordering of the data which they were designed to order would in itself produce scientific generalizations of the greatest importance. In actual fact, this did not happen at all. The techniques were means of developing variables that had certain desirable properties. Possibly fundamental laws may be discovered through the use of these variables.

Relationship of Research to Practical Problems: Action Research

Scientific research in education, like research in any other field of endeavor, is not necessarily directed toward the solution of some immediate and pressing practical problem. Many such problems can be solved only after a large body of knowledge has been accumulated, at which point they become researchable. We have emphasized here that the long-term program of research is the one that ultimately

provides educators with the power to predict and control events. The educator has felt impatient with the research worker who takes this position, and his impatience is shared by practical men in other fields who desire to introduce improvements but who are not willing to wait for research to tell them what to do. During World War II this issue became a serious one, and some of our European allies tried out schemes of employing scientists to recommend and institute innovations without recourse to full-scale research. The scientists were required to formulate plans in terms of the best knowledge available, and after the plans were placed in action some attempt was made to obtain evidence concerning their worth. They were not expected to undertake complete and thorough inquiries concerning the value of their innovations. The procedure was designed to expedite change and development and was probably an influential one. It was the father of what is now termed *operations research* and paralleled closely what is termed *action research* in education.

The advocates of action research in education also commonly stress the idea that those who are to be involved in educational change should be the main participants in the research process. Thus, it is suggested that teachers in schools should attempt to solve their own educational problems by establishing action research programs. It is claimed that such participation will ensure that whatever changes are demonstrated to be desirable will be brought about by the school personnel involved. In this way resistance to change will be at least partly broken down.

A perusal of the literature indicates to the writer that much of what is advocated as action research is nothing more than good management. Any modern book on management will suggest that when problems arise, an attempt should be made to draw up a list of alternative solutions, then to collect data to determine which one of these proposed solutions is best. This is good management practice but, in education, it goes by the name of "action research." It is contrasted with scientific research in that it does not, except by chance, build up a body of organized scientific knowledge. It may be highly effective in solving local problems. It may also serve the very important function of stimulating thought as well as change. It may produce useful knowledge. While there is much controversy in modern educational literature concerning the value of this approach to educational

problems, the practices of action research are so obviously consistent with good management practices that they are here to stay even though they may be given a different name a few years from now. Despite all these merits, the writer cannot see that action research and scientific research, as conceived here, have much in common.

An Overview of the Content of Educational Research

When the range of phenomena that the educational researcher may study is contemplated, it is realized that he should approach his task with considerable humility, recognizing that only a few of the foundation stones of an organized science can be laid within his lifetime. The hopes of those who first started research work in this field some fifty years ago have not yet been realized and will not be realized for a long time to come. Research did not provide a rapid revolution in which knowledge of the soundness of educational procedures was to replace prejudice and tradition; for as research workers began their inquiries, they soon realized the immensity of the task that lay before them. The areas of educational research may be likened to areas on a map that have been roughly circumscribed to indicate gross differences in terrain. Some penetration has perhaps been made within the borders of these areas, but most of them remain unexplored. Explorers of the future will provide broad knowledge of these general areas, and then must come the developers who will exploit the resources that each domain has to offer. The boundaries that have been set up are artificial, for each one of the areas of educational research fuses into the others. The criterion of relevance for study is purely a matter of whether an area has impact, direct or indirect, on the development of the child.

Summary

1. Educational research was initiated little more than half a century ago and set itself the ambitious task of producing a rapid revolution in educational practices. Time has shown that research in an area develops only slowly, and that decades may pass before suitable research methodologies are developed.

2. Since the central focus of educational research is the development of the pupil, and particularly insofar as this is produced by the practices of the school, it is hardly surprising that developmental studies constitute a major area of educational research.

3. Developmental studies have shown an evolution in the theoretical position from which they have been undertaken. Earlier studies were often dominated by the standpoint of the maturationist who sees development as largely the product of inner forces. Current research, in contrast, emphasizes the role of learning in the development process, and studies have become more and more concerned with the acquisition of new behavior as it is brought about by the school. The wealth of theoretical viewpoints from which the research worker may draw his ideas ensures that research in the area can be thoroughly theory oriented.

4. Since those who work in the field of curriculum emphasize the need for curriculum *theory* as an essential element of curriculum development, the stage is well set for programmatic curriculum research that is firmly rooted in clearly stated theories. This is a relatively recent development, for curriculum research of the past has tended to be a fact-finding enterprise. Many techniques are now available that permit research on important curriculum problems, the application of which has hardly yet been explored.

5. The influence of sociological and economic conditions on education is an important field of inquiry. There are, of course, real difficulties in developing generalizations that have any wide applicability in this area of investigation, but some that have been made are extremely important in their practical consequences.

6. The design and engineering of equipment for schools is a relatively new field of educational research and development. Current attempts to develop devices that provide effective learning situations represent a most important innovation. Such devices can play an important role in the acquisition of a wide range of thinking skills.

7. Educational research conducted within institutions for solving local problems usually contributes little to the development of a body of knowledge of broad significance to education as a whole. The pressure to solve one local problem after another prevents the development of a broad program consisting of a series of related studies. Institutional educational research does not have to be of this character.

8. This book does not attempt to discuss philosophical or historical research. The methodologies involved in both of these areas are so specialized that advice concerning them should be given by professional philosophers or professional historians.

9. Educational research may be carried out at a number of different levels of complexity, which range from the solving of local problems by well-known techniques to research that results in new theoretical developments of wide significance.

Selecting the Problem | 4

Allowing Time for Planning

The planning of research is commonly thought of by the novice as an initial stage that is quickly passed and followed by the more elaborate and prolonged stage of collecting data. Many of the weaknesses in current educational research are attributable to this viewpoint, which is fundamentally unsound. The fact is that the major effort in the undertaking of research should be devoted to the planning stage, which may include not only a careful formulation of the problem as outlined in the second chapter but also some preliminary data-collecting activities. Once a research has been well planned and the techniques have been given a preliminary trial to make sure that they are feasible methods of attacking the problem at hand, the actual execution of the research is a simple and mechanical matter, which requires more patience than brilliance. Weeks may go into planning an experiment that may be completed in a single day. The hard part of all research is the planning stage, which is the thinking stage.

Conant (1946) has pointed out that there is no simple formula to

help the researcher in the most crucial stage of the development of research, which is the stage of developing hypotheses. The would-be researcher must recognize that much brilliant work owes its brilliance to the significance of the hypothesis that is tested. Some researches, of course, are brilliant because of the unusual and ingenious way in which the hypothesis is tested. The unfortunate fact is that most research conducted in education is nothing short of drab in both conception and development. If this chapter can do anything to reduce just a little the drabness in research in education, it will have achieved its purpose.

There can be no doubt that individual researchers differ greatly in their sensitivity to problems. Guilford in his studies of creative talent has found that such an ability appears to be quite independent of other aspects of talent. This is in keeping with the experience of research administrators, who know that some researchers seem quite unable to locate and identify problems even though they may do a workmanlike job in solving problems assigned to them. No particularly useful advice can be given concerning ways of developing problem sensitivity, but one suspects that experience is an important factor. This is not particularly useful advice to the graduate student who is beginning his research career. Attendance at and participation in research seminars may help to sharpen the student's ability to discriminate between researchable and nonresearchable problems. Seminars may also bring the student into contact with researchers who are highly sensitive to the existence of researchable problems, and this experience may help him in developing his ability. While it is possible in this chapter to offer the student some help in identifying problems, little is likely to be achieved in the direction of making the student more sensitive to those that are researchable.

The student of educational phenomena should embark on research with full respect for the complexity of the phenomena with which he is faced. In this regard, much educational research lacks the humility that is essential if matters of importance are to be discovered. The writer can recall instances where persons who should have known better have approached educational problems as if they were of the complexity of a party game. On one occasion, a professor in charge of an educational research project remarked to him, "This year we are going to settle the problem of measuring teacher personality so

that next year we can move on to other matters." From our present perspective, it is quite clear that one hundred years from this time scientists will still be attempting to measure some aspects of teacher personality. The student may reflect that even in an area where problems seem so simple, relatively speaking, as in rote learning, investigation soon reveals that the phenomena studied are of enormous complexity. How much more complex must be the phenomena that take place in the classroom, or even those occurring in miniature educational situations set up for research purposes. Herein lies the central difficulty in identifying researchable problems.

Much of the deceptive simplicity of educational phenomena stems from the fact that considerable progress has been made in predicting success and failure in different types of curricula. The fact that persons with little research experience have been able to develop tests for accomplishing just this adds to the deception. Such straightforward relationships represent fortuitous rather than typical circumstances. The discovery of clear relationships beyond this point is a much more difficult matter. Sometimes even years of research on a single problem may yield but small returns.

The Acceptability of a Research Project in Relation to the Social Milieu in Which It Is Undertaken

When he embarks on any research, the student of education should recognize that, while he may wish to pursue his own whims and fancies, he is not entirely free to do so, for the acceptability of his product to others may determine whether he does or does not obtain a degree. Most scientists work under a similar restriction. The industrial scientist needs to recognize the goals of the concern for which he works, and at least to some small extent he must modify his own goals to make them compatible. In government service, it is necessary to realize that only certain types of research projects can survive over the years, and if a scientist embarks on a long-range program he should also have other short-term programs that will yield more immediate results of practical value.

This problem is an old one. Leonardo da Vinci found it necessary to spend much of his time devising instruments of war so that his patron would permit him to engage in scientific research. The system of patronage of the last century always required the scientist to

modify at least a part of his pursuits to conform to the desires of his sponsors.

There is, of course, much evil in the fact that in most situations the scientist must take cognizance of outside forces that can influence the acceptance or rejection of his work. It might even be argued that many notable discoveries have been made in flagrant opposition to current ideas. The work of Galileo in the face of social opposition is familiar to every reader. While the older scientist who is well established can perhaps afford to be insensitive to the social milieu within which his work takes place, the younger scientist finds need of this sensitivity, if only to reach the point where he can afford to present highly novel ideas that oppose current concepts. Here the writer is not endorsing the idea that the graduate student should pursue a line of thought thoroughly acceptable to his elders. He is only pointing out that the faculty of a school of education represents what might be called a subculture, and that, like other subcultures, it will accept some ideas much more easily than others. A graduate student can be expected to select a school where his ideas and those of the faculty display some degree of harmony. Related to this is a matter that must be considered now.

The opinion of the present writer, which is shared by many of his colleagues, is that it is a mistake for the researcher to orient his work in relation to some social issue about which he has deep personal convictions. While such convictions may stem from the most desirable and highly esteemed values, there are reasons why they form an unsound basis for research. In the first place, they usually lead the graduate student to attempt a problem that is way beyond his capabilities, and one that is often beyond the scope of available techniques to solve. The common trait of overambitiousness seen in so many doctoral studies is most often an outgrowth of the individual's own personal values, and that he seeks evidence that will support some private belief. Much wiser would such an individual be if he developed a research project as an outgrowth of another's systematically developed program.

A further difficulty also stems from the researcher's personal involvement in issues that pertain to his research. This is the difficulty of maintaining an objective attitude in the analysis and interpretation of data. Darwin noted that he found a tendency in himself to forget

those facts that were not in accord with his theory. Even more prone to forget such disagreeable facts is the person who has the deepest beliefs about the value to be attached to a particular viewpoint.

Finding Problems

It is difficult to supply definitive ideas concerning how the student should obtain ideas for his research. Part of the trouble arises from the fact that we know little as yet concerning the usefulness of various techniques for this purpose, and the scientist who is asked to say how he finds his ideas is usually quite unable to give a definitive answer. Certain procedures that the student may possibly find useful can be suggested, but these have not been validated.

One method of deriving research ideas is to read articles published in the current literature and to consider how the techniques and ideas discussed might be applied to the solution of other problems. The adaptation of techniques to the solution of new problems is a profitable and worthwhile enterprise in which many scientists engage. Indeed many such enterprises may be judged to have high originality.

A second approach is to identify in the literature studies that would have had merit except for some central defect that makes it impossible to draw conclusions from the findings. This is not to be looked down upon as an activity, for it often yields results of great importance.

A third procedure is to refer to the discussion sections of technical papers. In most such papers this section usually is the final paragraph, and it presents the author's reflections concerning the significance of the results and what type of investigation should be undertaken as a follow-up. Such suggestions appear in considerable numbers in the concluding sections of technical reports. Many, of course, present only ideas rather than practical suggestions, and many are beyond the realms of usability at the present. Nevertheless, fruitful ideas may still be found in quantity.

All of these procedures involve a review of the literature. Since the latter is not the simple matter that it may seem to be, some comment on this activity must now be made. The locating of published material requires well-developed techniques for using a library, and this book assumes that the student has those skills. If he does not, he is referred to an excellent book by Alexander and Burke

(1950) designed specifically for providing that type of training. Their volume, while excellent for its purpose, is limited to the mechanical aspects of using a library. It does not concern itself with the more subtle subject of how information derived from a library should be used. A person may locate all of the relevant references and still fail to derive from them the relevant information.

Unfortunately, a tradition has grown up that a "review of literature" is a low-level task, which can be undertaken by the student who is not very advanced. Many, of course, would disagree with this view, as is evident in the fact that the chapters in each *Review of Educational Research* are usually written by the senior members of the profession. In a similar spirit, the *Annual Review of Psychology* is written for the most part by persons who have had considerable experience in the fields they cover. A first-class review of the literature requires the maturity of viewpoint that comes from years of study and research. The student of education who has had a brief experience in graduate school cannot be expected to have the intellectual maturity to prepare a thoroughgoing review of research in an area of education in which he is interested, but the experience of making the review can be a worthwhile one, and with a few precautions much can be done to give it a professional and polished appearance.

Advice commonly given in making a review is to start by preparing a fairly complete list of references, but this is poor counsel, and any person who has engaged extensively in such work will know it. The first thing that a would-be reviewer must do is to familiarize himself with the issues and problems of the field. Until he has done this, he cannot possibly know what are and what are not relevant contributions.

Familiarization with the issues and problems of a field can be accomplished by reading articles that treat of this matter. In most areas, such articles exist, but these should not be confused with the type of article that is noninterpretative. For example, most chapters in the *Review of Educational Research* are noninterpretative and are mere listings of studies that have been made in the previous three-year period.

Critical review-type articles serve the purpose of indicating to the student what are the central issues to be taken into account in his own reading and review of the literature. He would also do well to discuss his early impressions of the literature with some professional

person who is thoroughly familiar with the area. He will then be ready to begin work on his own review.

Through his preliminary reading, the student will have located some of the major references. These should be consulted next. At this stage a good plan is to enter the title of each reference at the top of a five-by-eight-inch card and to use the remainder of the sheet for summarizing the article with particular reference to the light it throws on major issues. A small section at the bottom of the card may be reserved for critical comments and further hypotheses suggested by the author of the study.

Additional references will be found in each additional article that is reviewed, and thus most of the significant sources will be obtained. At this point, the reader may ask why it has not been suggested that he prepare a comprehensive list of references from a source such as the *Education Index*, which lists every article and publication that has any relevance at all to educational problems. The answer is that in such a source the classification of references is necessarily very crude and often depends more on the title of the article than on its content. Such comprehensive lists of publications may supply a rough check on what is available, but they cannot provide a basis for a critical review. However, such lists do permit a superficial independent check of the completeness of the references obtained from published articles.

In the development of his review of the chosen field, the student will have opportunity of discovering a problem that he can use as a basis for his research. If such a problem has been found, then the review can be written in terms of the relevance of the various studies to it. Until such a problem is found, it may be well to postpone the final integration of the material into a single review, which will usually constitute the first chapter of the thesis or dissertation.

The reader should pause at this point and link the considerations of this chapter with the earlier discussion of the need for conducting research within a framework of theory. *The review of the literature, if it is conducted in the way described here, should provide an overview of the current framework of theory in the area in which it is proposed to undertake an investigation. The student may be expected to abstract from his review of the literature a theory in terms of which he plans to work. A minimum requirement should be that he draw up a statement covering the essential features of the theory, but*

preferably he should be more ambitious and draft the theory as a set of postulates. He should then show how his hypotheses represent a series of deductions from these postulates. This he will find to be a worthwhile exercise in clear thinking.

SOME POINTS ON THE EVALUATION OF RESEARCH STUDIES

In previous sections it has been emphasized that the person who embarks on a research should be thoroughly familiar with previous work in the area and should attempt to organize this knowledge into a theory that forms the basis for future work. It is perfectly obvious that all that has been written and published in an area cannot be given equal weight in such a review. Indeed, much of the available material may be disregarded when the final summary is made. Up to this point no advice has been given concerning the evaluation of published studies as a basis for retaining or discarding material.

In a sense this entire book is concerned with the evaluation of research, for any attempt to teach the student something about methodologies, if it is successful, must help him to judge between adequate and inadequate research. A summary of much of this book would be a list of procedures that some scientists have found useful for advancing knowledge, plus a list of common errors that are likely to render studies useless or to reduce their value greatly. The chapter summaries collectively do this in a fairly concentrated form. However, the writer feels that there would be merit in presenting at this stage a rather brief section on the topic of evaluating research studies. If the student is developing a research project at the same time as he is taking a course on research methodology, he will need to perform these evaluative functions long before he has studied problems of research development and design. Nevertheless, a mature ability to evaluate research will require much more extended study than has been undertaken up to this point.

Evaluating the Problem

There are several ways in which it is possible to evaluate the problem that forms the focus of a research, and most of these have already been stated in one way or another. First, in reviewing a research study and evaluating its worth, the reader may well ask, "Is

the problem clearly stated?" If the problem is not clearly stated, it is quite evident that the research cannot make any significant contribution to knowledge. The statement of the problem should be found in the early paragraphs of the research report and should be preceded by only those materials that are necessary for a full understanding of the problem.

Second, it may be asked, "Is the relation of the problem to previous work in the area clearly stated?" This is a point on which it is a little more difficult to evaluate a study than is at first evident. Sometimes an expert reading an article may see very clearly the relationship between the problem and previous work, while a novice in the field may not. This is in part due to the fact that many technical journals frown on long introductions to articles, although such introductions are really necessary for the novice to see clearly how the research is related to previous work. Such journals cut down on introductory materials merely as a space-saving device. It is unfortunate that this is necessary, since as a result such articles can be fully understood only by those who have already read extensively in the field.

Third, the reader must ask himself, "Is the problem broad enough to provide a study of real significance?" Many researches that are conducted in the educational field pertain to local problems, and their results have no significance whatsoever for other areas. The narrowness of a problem is a particular problem in experimental studies where the results apply only under conditions of a highly specific type.

Perhaps one further question may be asked in the process of evaluating the problem—"Does the solution to this problem pave the way for the development of further knowledge?" A problem should not lead up a dead-end street, but rather should it be an avenue that opens up new territory. A research designed to solve some local problem within a school system is unlikely to meet this criterion. Such a study will probably, though not necessarily, be given little weight in a review of the literature, and much less weight than one that attempts to solve a problem of rather broad significance.

Evaluating the Procedure

The first point to note in the evaluation of the procedure adopted in a study that is being appraised is that the description of it should be sufficiently precise to enable another person to reproduce the study.

If essential aspects of the procedure are not described, it is not possible to verify the results by undertaking a similar study. If the procedure is not adequately described, it is also not possible to evaluate its merits, since the aspects omitted from the description may have been quite inadequate even though the remainder may have been adequate. Of course complete description is rarely if ever possible, as will be pointed out later in this volume.

While the procedure is often well described, it is common to find that the nature of the population to whom the procedure was applied is not even mentioned. This is an important omission, for if a study were undertaken using as subjects state patrolmen, it would hardly be reasonable to apply the results to, say, elderly spinster teachers.

Then the reader must ask himself, "Are the procedures adopted in the study clearly related to the solution of the problem?" Much more is to be said about this point in later chapters. For the present, we wish only to warn the student that many published researches never amount to anything simply because the procedures adopted could not accomplish what they were supposed to accomplish. A common error in research is to use a psychological test without being sure that the test is valid for the purpose for which it is being used. If the procedures and techniques are not suitable for solving the problem at hand, it is evident that the study can be of little value. Sometimes a research worker recognizes the fact that the use of certain procedures involves certain assumptions, and if he is wise he states those assumptions. He may make the assumption, for example, that a measure of spelling derived from a published objective test of the skill provides a true indication of the pupil's habitual skill in spelling as it is manifested in his schoolwork. The research worker, or the reader of the account of the research, may be able to point to other studies that supply evidence in support of or contrary to this assumption. This evidence may in turn be used as an aid in the full evaluation of the study.

The Design of the Study and the Adequacy of the Analysis

Any research that is reviewed must be appraised partly in terms of the extent to which it was adequately designed. Problems of design are considered at some length in later chapters, but a brief discussion of them is in order here. Design has two aspects. First,

there is the matter of whether it permits the collection of the evidence that is necessary to solve the problem. Second, there is the matter of whether it is efficient—that is to say, whether it permits the collection of the maximum amount of information with the least amount of effort. The evaluation of designs is quite a technical matter, and there are no simple and straightforward methods for it that can be given in a few paragraphs. At this stage the student should look only for gross flaws. It is surprising how many such flaws do occur in published studies.

The Evaluation of the Results and Conclusions

No study is of any real value unless the results are clearly stated in a form that permits one to know precisely what was found. If one cannot determine just what was found, or if there is ambiguity concerning it, the worth of the study must be questioned. Sometimes one does not know just what was found because some of the data appear to have been omitted. The writer can recall a study in which a series of tests was administered to five hundred cases, but the results were reported for only about four hundred, which left one wondering what happened to the other one hundred cases. Were the data for them lost? Were they cases that provided data inconsistent with what the researcher was trying to demonstrate? One can only hope that the latter was not the case, but for all the information given it might have been. Many ambiguities in the presentation of results are due to the fact that the research worker does not indicate exactly what happened to the data that he collected.

If the results are clearly presented, then the reviewer should examine the conclusions and determine whether they follow from the results. Then he should determine whether the conclusions are consistent with those of other studies in the same general area. If they are inconsistent, he should see whether the author of the research has a reasonable explanation for the inconsistency.

Some Additional Criteria for Evaluating Published Research

There are certain other criteria that the student may wish to use in evaluating research. Place of publication is certainly an important cue. If a research study has been published in a well-established technical journal with a high reputation for the quality of its con-

tributions, it is probable that it has been reviewed by competent experts in the area and has been found to be a study of quality. Of course, the experts are sometimes wrong. Articles published in more obscure sources and in journals that require the author to pay all publication costs are much less choosy in what they accept. It is hardly surprising that the sources of free publication are able to select the best contributions.

Authorship is a somewhat controversial basis for assessing research publications. Books and articles in technical fields are usually evaluated for possible publication without the name of the author being indicated on the manuscript. The novice in research, and perhaps the mature research worker too, may well feel that the work of the renowned expert can be expected to be outstanding. Again, one must emphasize that this is not always so.

The Effect of Selective Publication on Reported Results

In reviewing the literature, the reviewer should be aware of the factor of selective publication. The writer believes that this phenomenon is well illustrated by the studies on the merits of progressive education that appeared during the years 1920 to 1940. Those that appeared in the literature during this period were designed mainly for the purpose of comparing the relative merits of what was termed "progressive education" and what was termed "traditional education." Their general procedure was to measure the achievement of two matched groups that had been exposed to these two types of curricula. Some years ago the writer reviewed this literature, and noted that the small studies almost always favored the progressive curriculum while the larger studies were much less favorable. The best explanation for this seems to be that strong feelings are involved in this area, and most of those who undertake studies in it are motivated by the desire to demonstrate that progressive education is superior to the traditional approach. On this account, one suspects that, when a small study is undertaken and produces results that are unfavorable to the cause of progressive education, the tendency is simply not to publish the study. In contrast, when a small study produces favorable results that confirm the research worker's opinion, considerable effort may be made to find a publisher. However, when a large and extensive study is involved, the research worker is likely to have in-

vested so much time and effort that he cannot afford not to publish, even when the results run counter to his previous opinions.

The reader can see that selective publication is likely to result in the data providing a biased idea of the extent to which a particular point of view can be supported by evidence. A reviewer of technical literature should keep this possibility in mind.

DESIRABLE CHARACTERISTICS OF THE PROBLEM

The problem that is eventually isolated may be stated in terms of a question for which the proposed research is designed to obtain an answer. Sometimes the question to be answered is referred to as a hypothesis. Sometimes in this book it has been called a deduction from a postulate. Certain criteria may be suggested for judging the merits of hypotheses, and these need to be discussed further at this point. It will be assumed in this discussion that the hypothesis is firmly rooted in a framework of theory, and hence this particular criterion will not be discussed here at further length.

Hypotheses selected for research should be testable. One of the commonest sources of difficulty for the graduate student who embarks on a dissertation is the selection of a hypothesis that is not really testable. The same difficulty is also apparent in the researches of some of the more mature members of the educational profession. For example, one educator selected for his research the hypothesis that secondary schoolteachers did not know enough algebra to teach their pupils competently. This is not really a scientific problem but nonetheless one of some interest. He proceeded to test this untestable hypothesis by administering an algebra examination he had devised to a group of secondary schoolteachers. Since the questions in his test gave the appearance of having been devised to confuse, it is hardly surprising that most of the teachers achieved a very low score. His conclusion was that the teachers did not know enough algebra to teach with competency, which was just a reiteration of the opinion he had held in the first place. The data really provided no genuine information to endorse or reject the conclusion. Since he wanted to "prove a point," he had done this by devising a test that measured the essential knowledge of the teacher, according to his own judgment (and few might agree with him). What was needed in order

to make his hypothesis a testable one was a prior study establishing what mathematical knowledge was, and what was not, essential or desirable in an algebra teacher.

This point should be emphasized because some of the most interesting and important hypotheses are quite untestable at this time. It is important to learn this lesson, because a common way of attempting to select problems for a program of educational research is to start by listing the problems that are in most urgent need of attack from the practical standpoint. The author has been associated with several projects set up in this way and has invariably protested the use of this procedure, but the result has always been the same. The research program has bogged down in a swamp of untestable hypotheses. While the researcher may begin his thinking with some focal point in education where answers are urgently needed to important questions, he should start by struggling to find a contact point between available organized knowledge and the problem with which he is confronted. If such a contact point does not exist, the researcher must assume that he is attempting to operate in an area that, because of its isolation from organized knowledge, is likely to yield untestable hypotheses.

Hypotheses should state relationships between variables. A well-developed hypothesis that meets satisfactory standards should state an expected relationship between variables. Unless hypotheses can be stated in this form, they have not reached the point where they are appropriate as a basis for research. A hypothesis such as "Children who attend Sunday school show greater moral growth than children who do not" is not testable, because the term "greater moral growth" does not refer to a variable that is measurable at the present time or likely to be measured in the near future. On the other hand, a hypothesis such as "Teachers who manifest aggression in the classroom have pupils who also manifest aggression" refers to a variable, aggression, that can be measured through such procedures as counting the number of specific types of aggressive incidents that occur. However, the reader should recognize the fact that it is often necessary to use quite indirect means of measurement. The latter is true of all sciences. The physicist measures the amount of various elements in the sun by studying the spectrum of its light. The psychologist may attempt to measure emotional disturbance through the

response of the individual to an ink blot. Although hypotheses should state relationships between variables, it does not mean that these variables have to be measured by any direct method, although any indirect measurement should be based on a clear-cut rationale.

Hypotheses should be limited in scope. A common error of the graduate student of education in planning research is to develop hypotheses of global significance. It is perhaps natural for the beginning research worker to be overambitious in his initial efforts, partly because of his earnestness and partly because it takes maturity of viewpoint to realize how little can be accomplished in a lifetime. The more mature research worker is likely to choose hypotheses narrower in scope and therefore more testable. The student should seek hypotheses that are relatively simple to test and yet highly significant. He should try to bring order into a very limited corner of the universe—but it should be an important corner.

Hypotheses should be consistent with most known facts. Any hypothesis formulated as a basis for research must be consistent with a substantial body of established fact. It is too much to expect that it be consistent with *all* established facts because in so many areas the facts themselves appear to be inconsistent with one another. For example, in the area of vision, it is known that single nerve fibers cannot carry more than one type of impulse; yet there do not seem to be a sufficient number of nerve cells in the retina to make up three distinct mechanisms for the three primary colors. No theory of vision has been able to resolve all of the apparently inconsistent facts, and almost any hypothesis formulated is likely to be consistent with some of the facts and inconsistent with others.

Hypotheses should be stated as far as possible in simple terms. This is desirable in part to permit the meaning to become clear to others, but it is also desirable since in order for a hypothesis to be testable, it must be stated in relatively simple terms. The simplicity of the statement has nothing to do with its significance. Some of the most important hypotheses ever tested have been such as could be explained to an average child in junior high school. It is the simple truths tentatively formulated as hypotheses that form the fundamental cornerstones of science. For example, Pasteur's hypothesis that life would not be spontaneously generated from organic matter if all living matter were first destroyed is an easily understood concept, yet

it is one that deals with a matter of fundamental importance. Newton's hypothesis that a body continues in uniform motion unless acted upon by a force is a simple one, yet it became the cornerstone of physics.

Hypotheses should be simple from another point of view. They should avoid the use of vague constructs, however popular these may happen to be in current educational thought. It is quite useless to formulate a hypothesis such as, "The adjustment of the pupil to the classroom situation will depend upon the total classroom situation." Such a hypothesis includes several vague concepts, one of which is "the total classroom situation." To say that an event depends upon everything else that is happening fails to do what the scientist has to do, namely isolate a few aspects of his environment that have special relevance as factors in the production of the phenomenon in which he is interested. The specification of these characteristics must be undertaken in the formulation of a clear and simple and important hypothesis.

The hypothesis selected should be amenable to testing within a reasonable time. The student of education is too often excessively ambitious when he first seeks to undertake research. This is usually a result of the fact that he is in close contact with the pressing problems of education. He is frustrated by being perpetually confronted with problems that must be solved before major advances can be made, and to overcome his feeling of personal frustration he sets himself the goal of solving one of these major problems. Yet the fact is that nearly all such problems cannot be solved for a long time to come. They are mainly problems of immense difficulty, which cannot be profitably studied because the essential techniques for attacking them are not available. This is well illustrated by the numerous graduate students of education who suggest each year that they develop a doctoral dissertation in the field of teacher effectiveness. The common proposal is that a study be made of personality traits as related to teacher effectiveness. Such studies assume that measures of relevant teacher traits are available—and of course they are not. These studies also assume that the effectiveness of the teacher in achieving various kinds of significant objectives is known, and that the growth of the pupils with respect to these objectives can be measured. It is almost certainly true that most of such achievements

cannot be measured at the present time and that no means will be found to measure them for a long time to come. Most studies of teacher effectiveness that the graduate student is likely to consider are impractical because the techniques for carrying them out simply are not available and cannot be developed rapidly.

The student should be warned against doing what is commonly done when the would-be researcher finds that techniques are not available for the study of a particular problem—that is, using what are often hopelessly inadequate techniques. For example, many who have wished to study personality characteristics of teachers related to their effectiveness have ultimately settled for studies involving the correlation of ratings of teacher effectiveness with ratings of personality characteristics. Such activity can be described only as pseudo research. It bears a relation to well-conducted research in that it involves the statement of a hypothesis and the collection of data, but the data have only a superficial relationship to the testing of hypothesis. The serious research worker would find it hard to accept the belief that actual teacher effectiveness in achieving a particular objective is related, except to a slight extent, to ratings of effectiveness, produced by an observer, for the judgments of an observer are likely to be very erroneous. Similar doubts may be expressed about the validity of ratings of the teacher's personality characteristics. What such research produces is correlations between one obscure variable and another, and in this obscurity little light can be discerned. If, on the other hand, it were possible to measure the true amount of learning produced by each of several teachers working with comparable classes, it would be of the greatest interest to determine the relationship between the personality characteristics of the teachers and the amount of learning produced. In this case the data would be meaningful and would not represent the type of worthless substitute for meaningful data so commonly introduced into educational research. Educational literature is full of examples of studies in which a student's enthusiasm for a problem has blinded him to the weaknesses of the techniques through which he has tried to study it.

Sometimes, in order to test a given hypothesis, it is necessary to test a hypothesis related to it by some rather remote channel of reasoning. An example from the physical sciences illustrates this point more clearly than examples from education. In order to test a hypoth-

esis concerning the nature of the chemical and physical processes that result in the release of the sun's energy, it is necessary to make a spectral analysis of the light coming from the sun. From this analysis, it is possible to make inferences concerning the conditions that resulted in the production of the light.

Indirect Versus Direct Approaches

A common error in educational research results from the researcher attempting a too direct attack on his problem. For example, generations of educational researchers have attempted to appraise the level of motivation of students and others by asking persons directly to indicate how well motivated they are, or by asking observers to indicate how well motivated these persons appear to be. Such approaches are well known to be quite futile, for the work that Freud initiated, which has now been pursued for nearly a century, has shown quite clearly that motivation is a matter of which the individual is largely unaware and that he has the greatest difficulty in explaining. The same clinical studies indicate that direct observation of behavior as it occurs on a day-to-day basis reveals little concerning motivation. Motivation is, as it is commonly said, a hidden variable that cannot be observed directly and cannot be assessed directly. It is the indirect approaches to motivation, such as those of the clinician or those that began with the work of H.A. Murray and his associates in the 1930's, that have yielded the little knowledge we now have concerning it.

Indirect approaches to problems are typical of all branches of science. The realization that the laws of falling bodies could be studied best, not by studying free-falling bodies but by studying such artificial situations as objects moving down inclined planes, opened an entirely new era in physical experimentation. The study of human genetics has been made possible through studies of the microscopic structure of plant cells. The development of radioactive materials has made it possible to investigate human metabolic processes that have defied any direct approach. Much scientific knowledge has to be acquired by indirect methods. Even the practical problem of measuring the diameter of the earth does not lend itself to the direct approach, which would involve the stretching of a measuring tape around its circumference. All knowledge about the atom and its supposed structure is acquired by extremely indirect methods, where the measure-

ments made are connected only remotely with atomic phenomena and where the conclusions involve a long chain of supposed events.

The directness of approach of many who work in educational research is not too different from the approach of the physical scientist in the Middle Ages who wished to solve problems of converting lead to gold by simple and direct means. Part of the reason for this is that our ingenuity has not led us yet to useful indirect methods of attack on more important questions that are likely to arise, and thus we tend to keep hammering away with direct approaches, which mostly have no value at all. However, we can take a few steps in the right direction. Even an opinion survey at times can avoid to advantage a direct question concerning a problem. For example, a person may be unwilling to admit what he pays his servant, but he may be quite willing to state, without embarrassment, what he believes to be the prevailing wage in the community for that kind of work.

Sometimes the indirect approach to problems involves the conduct of a study in a laboratory situation rather than in a real-life setting. Many problems of reading have been attacked successfully in this way, and subsequent classroom studies have validated the results. There are advantages in a direct approach whenever it is likely to yield results, but the student who finds that only an indirect avenue is open to him should not feel discouraged. He should remember that some of the most important discoveries of science were made by a similar means.

Research in administration is an area in which a direct approach is often not feasible but in which indirect attacks on the problem may be highly productive. The writer can recall the suggestion of a student interested in the question of how information was passed around in a large board of education building. The suggestion was the simple one of keeping a record of who called who on the telephone within the building. There was no intention of keeping a record of what was said, for the purpose was only to draw up a diagram rather, like a sociogram, that would indicate the channels through which information passed during the course of daily business. Of course there has also been much laboratory work conducted on the effect of various administrative practices on the morale of groups, and these studies are being slowly extended into the field of real administration. The choice of level of reality of a study, its directness

or indirectness, is determined by a multiplicity of factors, including the amount that is already known about the phenomena.

The Data Language

The behavioral sciences, as they have developed in recent years, have shown a tendency to develop their own language and to discuss events in terms quite different from those used in everyday speech. Those engaged in research have become increasingly aware of the importance of selecting a suitable language with which to describe and discuss the events in which they are interested. The language selected is referred to as the *data language*. Since the data language of a study is used, in the first place, in the statement of the hypothesis to be tested, some problems related to the selection of a data language must be discussed at this time. The problem to be investigated and the specific hypothesis to be tested must be stated in a language that is appropriate.

Consider the case of the researcher who is studying some aspect of the behavior of the teacher. Much of the data of any such study must be derived in some way from the movements of the teacher in the classroom or from disturbances that he produces in his physical surroundings, as when his vocal cords cause vibrations in the atmosphere. Now an educational researcher who described all of the movements of a teacher and all of the physical disturbances produced by his behavior in the physical environment during a one-hour period would not have a description of events that would be in the slightest respect meaningful to another research. Also a graph showing the decibel volume of noise produced by the teacher's larynx would not convey to most of those who inspected it just what had happened in this respect in the classroom. A language that described teacher behavior in terms of the physical properties of movement, direction, force, pitch, etc., would at the present time be an entirely inappropriate data language for any researcher who wished to study the classroom behavior of teachers. Now it is quite conceivable that a time may come when such a data language may be appropriate and hence may be meaningful to other researchers, but until that time comes it must be considered as inappropriate.

The development of a suitable data language for any program of research in the behavioral sciences must take cognizance of two facts. First, it must be based on the recognition that all behavioral events

that occur during a data-gathering procedure cannot be recorded as part of the data, for much more occurs than can ever be recorded. Second, what is recorded constitutes only certain aspects of behavior, and terms must be employed that are generally recognized as referring to the particular aspects of behavior that are abstracted. Third, the terms used must refer to objectively identifiable events, that is to say events that independent observers can identify. Thus, in describing the behavior of the teacher, the researcher may record the number of instances in which the teacher threatened the children with keeping them after school, giving them additional homework, discussing behavior problems with the parents, and similar specified matters. Such a category of teacher behavior is meaningful to other researchers who read about it, and one hopes that the researcher had planned his work in such a way that this category of behavior could be postulated to be highly relevant to the phenomenon that was the object of the research.

A data language should not contain unnecessary references to unobservables. For example, a researcher recorded the statement, "The teacher felt frustrated because he was unable to maintain order." This statement referred to the teacher's feeling of frustration, which could not be observed but only inferred from observables. The kind of inference that is implied in the researcher's statement should never be included in the data language; rather should the data language refer to the events on which the inference is based.

The data language used by a specific researcher in the behavioral sciences will depend to a considerable extent on the nature of his academic training and background. In the illustration used two paragraphs above, the data language was derived from the language of common speech. The key word was "threat," and this word was used in one of its common meanings, but the researcher is in no way bound to this type of language. Often categories of behavior that will be understood well only by scientists will be used. For example, one researcher found it convenient to classify teacher behavior as learning-oriented or threat-oriented. These terms refer to behavior that is hypothesized to contribute to the organized learning of the classroom and behavior that serves the purpose of defending the teacher's ego against some external threat. The latter is exemplified by a teacher whose principal tended to judge teachers in terms of the orderliness of the pupils in the classroom. For this reason, this

teacher spent almost the entire time exercising control over every movement of the pupils. Now the terms "learning-oriented" and "threat-oriented" in this context are familiar to the person with thorough training in psychology but quite meaningless to most other groups.

In selecting a data language, the scientist sometimes makes up new words or defines common words in new ways. Occasionally this is necessary, but such a procedure should be reduced to a minimum for several reasons. One is that it requires each reader to learn a new vocabulary before he can study the results of research. Few readers of the research will be inclined to do this, and still fewer will be willing to master the language to the point where it can be easily manipulated. Another problem arises from the fact that if familiar words are used in unfamiliar ways, the groundwork is laid for a long history of misunderstandings. Indeed whole areas of knowledge have been confused by this practice. A good example of this difficulty is presented by modern learning theory, where the term "reflex" has different meanings in the language of different theorists.

The data language usually refers to only a limited range of phenomena. The researcher should try to specify the range of phenomena to which his data language applies. This is important because often readers will assume that the researcher is referring to a much wider range of phenomena than was the intention. For example, a researcher may have defined the phrase "aggressive behavior" in terms of certain aspects of the behavior of the pupil in the classroom. A reader in going over the report may easily forget the way in which "aggressive behavior" was defined and assume that the term and the conclusions that refer to it apply not only to behavior in the classroom but also to behavior in other situations.

The data language may also refer to characteristics of the physical environment that are hypothesized to effect behavior in some way. These characteristics have been referred to in this book as stimulus variables, and they may include such varied features as intensity of lighting, number of books in the school library, number of pupils in the class, and the area of classroom space per pupil. The data language for discussing such variables is derived from the language of everyday speech and presents no special problem. However, there are many important characteristics of the school that cannot be so easily described. For example, what is commonly referred to as

"degree of permissiveness" is a characteristic of the school that is difficult to define in terms most persons would understand, and thus it becomes extremely important to tie this term down to observable events that can be recognized.

At a primitive level of scientific development, the data language usually consists of qualitative and descriptive terms. These terms become progressively more remotely related to the language of everyday usage as the science develops. At a more mature level, the terms of the language begin to refer to variables rather than to qualitative phenomena. In the ultimate stage of development, the language is in terms of mathematical symbols that refer only to measurable variables.

The Advantages of Breadth and Narrowness in Defining Problems

There are disadvantages in the definition of a problem in narrow terms, particularly in the early stages of exploration. Narrowness hampers the possibilities of an imaginative approach. This can be appreciated by presenting a concrete problem from a field other than education. The example here is one developed by John Arnold, who, in his classes on creative engineering, stresses the importance of defining problems at first in broad terms. He points out that in one of his classes some students embarked on the engineering problems of designing an improved automatic toaster. By stating the problem in this way, the possible ideas that could be incorporated in a plan of action were restricted. If the problem had been defined as that of *developing new methods of providing the consumer with toasted bread,* a wide range of new ideas would have become available for exploration. For example, one can conceive of the possibility of providing the consumer with ready-packaged toast. Industrial methods of large-scale toast-making could then be considered. There was also the possibility that some commercial substitute for toast might be developed. So long as the problem was that of *developing a better toaster,* these latter possibilities could not receive consideration. Now there is no question that ultimately a problem has to be narrowed before it can be worked upon, but this should not happen until opportunity has been provided to explore the problem on a wide base with the full play of imagination.

There are similar disadvantages attached to the early narrow definition of problems in the field of education. Thus, in the search for

a problem to work on in the field of mathematics, the researcher might well start by asking himself the question, "In what ways is it possible to improve the teaching of number operations?" rather than the question, "In what ways is it possible to improve the teaching of long division?" When the student begins to think in terms of the broad problem, he is free to identify some crucial aspect of the teaching of arithmetic the improvement of which would result in the improvement of the teaching of arithmetic in general. On the other hand, if the student thinks only in terms of teaching long division, the outcome of the resulting research is likely to be applicable only to the teaching of long division. The student should direct his thinking in such a way that the ultimate product of the research envisaged is a principle that has at least the possibility of being widely applicable.

It should be pointed out here that we are referring in this section to the early stages of developing research. As thinking progresses, it is necessary to consider more and more specific aspects of the problem.

Preliminary Explorations of the Problem

The selection of a problem for study is not usually undertaken in a single step, for it is commonly necessary to run a preliminary study before the decision is finally made. The need for such a preliminary study does not arise when the problem requires the conduct of a research closely similar to one that has already been done, for it is then known that the research can be undertaken. However, when the field of inquiry is relatively new and does not have available a set of well-developed techniques, a brief feasibility study must almost always be run. Such brief trial runs demonstrate whether it is practical to undertake the research, whether the professional techniques are sufficiently sensitive to measure differences that it is desired to measure, and whether one can obtain the necessary cooperation of others involved in the study. Negative results in any one of these directions may be sufficient to cause the researcher to change his problem.

A preliminary trial or pilot study also provides some indications of the availability of subjects, if human subjects are used, or of other needed materials. Certain studies may require specific population characteristics, and it is necessary to determine whether populations

having these characteristics actually exist. For example, one study required a comparison of the performance of children who did not like their teachers with children who did, and each one of these categories of children had to be divided into a bright group and a dull group. A preliminary study was needed to determine whether enough children existed who would admit not liking their teachers to make the study possible.

Preliminary trial runs involve not only the selection of a problem, but also the selection of some kind of design for the study. In practice the design of the trial run may be much simpler and less sophisticated than the design that is finally adopted. The trial run may provide much information that is needed for the final design.

Developing a Research Plan

A stage arrives in the development of every research at which it becomes desirable for the worker to arrange his ideas in order and write them down in the form of an experimental plan. A few experienced and sophisticated research workers may never actually write out such a plan, just as most experienced writers do not start by making an outline, but most research workers need a formal plan just as most writers need to make an outline. The student of education who is embarking on his first research enterprise will certainly need to develop and outline a research plan that will serve a number of different purposes.

First, the research plan helps him to organize his ideas in a form whereby it will be possible for him to look for flaws and inadequacies. Many research studies appear to offer excellent promise until the details are laid out in black and white. Only then are the difficulties of executing the study likely to become apparent.

Second, the research plan provides an inventory of what has to be done and what materials have to be collected as a preliminary step in the undertaking of the study.

Third, the research plan is a document that can be given to others for comment and criticism. Without such a plan it is difficult for the critic to provide a comprehensive review of the proposed study. Word-of-mouth methods of communicating the proposed study are more time-consuming and less efficient than that provided by a written plan.

A research plan should cover at least the items that are discussed

in the paragraphs that follow. Only a brief discussion is presented here, since many of the points are treated at greater length in the chapters that follow.

1. The problem. The plan should include a clear statement of the question or questions that the research is designed to answer. These are the hypotheses. The plan should also provide a concise account of the background of the problem and the theory on which it is based. The questions must be clearly and precisely stated. The statement of the problem must be complete, and it must be presented in a form that makes absolutely clear just what information must be obtained in order to solve the problem.

2. The method to be used in solving the problem. This section of the plan provides an over-all description of the approach that offers an avenue to the solution of the problem. Sometimes it is necessary to adopt methods that make special assumptions, and these should be explicitly stated in this section of the plan. For example, if the method involves the measurement of attitudes by means of verbal attitude scales, then it may be necessary to assume that verbal expressions of attitude are related to other expressions of attitude. In the latter case it might not be desirable to continue with the research unless evidence could be marshaled to show that the assumption was justified. Usually it is necessary to introduce assumptions about methods simply because direct attacks on the problem are not possible, and the indirect nature of the approach that must be taken introduces the need for assumptions.

3. Procedures and techniques. While the previous section describes the over-all approach to the problem, this part of the plan is concerned with the details of the techniques to be adopted. If interview methods are to be used, an account of the nature of the contemplated interview procedures should be given here, also whether the interview is to be structured and in what way, and the characteristics that the interviewer should possess for the purposes of the study. If tests are to be given, the conditions under which they are to be held should be specified, as well as the nature of the instruments that are to be used. This section is an appropriate place for describing apparatus to be used or to be built. If public records are to be consulted as sources of data, this fact should be recorded here. The details of the procedures and techniques would not be complete without an account of the

sample that is to be included in the study. A statement should be included indicating how the sample should be drawn and the universe from which it is drawn as well as its size.

4. The population of events to be studied. The population of events to be studied will depend upon the population to which the results of the study are to be generalized. If the results are to be generalized to all seventh-grade pupils in a certain school system, the research plan should state this fact. Since it probably will not be possible to include all seventh-grade pupils in the study, but only a sample, the research plan should state how the sample is to be identified. The method of identifying the sample should be such that generalization from the sample to the original population is feasible. If textbooks are the subject of the research, the population of textbooks to which the results are to be generalized must be specified and so too must the method of identifying the sample of textbooks to be studied.

5. Methods to be used in processing data. A research plan should indicate the statistical and other methods that are to be used for processing data. Such methods should not be left until the data have been collected. Many students have completed considerable work on a study, only to find that statistical techniques did not exist for answering the questions that were asked. This part of the plan should be reviewed by a person expert in the field of statistics, since such a specialist can often suggest changes that result in substantial savings of time and effort.

Summary

1. The planning stage of research is the critical stage and should not be hurried. Often it is the most time-consuming stage.

2. The student should be aware of the great complexity of the phenomena that are studied in educational research. Even the simplest of these are extraordinarily complex.

3. There are dangers in the student seeking to investigate some problem related to a social issue about which he has deep feelings. The difficulties of maintaining an objective view of the results under such conditions present serious problems.

4. One method of identifying a researchable problem is to become fully absorbed in the technical literature of the area of interest. Another is to identify previous studies that should be repeated with refinements.

5. A review of the literature is an important part of the activities that prepare the student to undertake research. He may well start this by studying articles by the outstanding scientists in the field. The review of the literature should provide the student with the framework of theory required for research.

6. The hypothesis that is to be selected for study should be testable and should be stated in terms of variables that can actually be measured. The hypothesis should also be limited in scope and consistent with known facts, and should represent a deduction from a theory that the student can identify.

7. Most problems that are researchable at all must be attacked by indirect means. Most of the questionnaire studies that abound in education are misguided efforts to obtain information by direct means.

8. The problem must be stated in a language that is appropriate. In other words, a suitable data language must be selected. The data language in terms of which a technical problem is stated is usually considerably different from the language of ordinary communication.

9. The fact that the student is able to identify a problem does not mean that he is ready to jump ahead and carry through the collection of data. Some preliminary studies are usually necessary in order to determine whether the proposed study is or is not feasible. Preliminary studies often indicate that a problem needs to be restated or modified before it can be considered researchable.

10. A research plan that outlines the essential features of the proposed research should be prepared. An important function of this plan is to provide an outline of the inquiry that others may review and criticize.

Some Problems for the Student

1. Students who have already identified research problems should present them to the class for criticism. Much can be learned by such a critical review. It is suggested that the problems be presented in the form of the research plan that has been outlined in this chapter. The student should be prepared to defend the theoretical position to which the proposed research relates.

2. Students who have not yet identified research problems should review the technical literature in the general areas in which it is proposed to work. The central theoretical concepts of the selected areas should be identified, and problems should then be outlined for criticism by others.

Measurement in Research | 5

Measurement and Science

It is difficult to conceive of a scientific approach to problems that does not involve the use of measurement. When measurement is involved, it is usual to say that *quantitative* methods are being used, and these methods are to be contrasted with *qualitative* methods, which do not involve the use of measurements. Quite obviously, much of importance can be learned by the use of qualitative methods, but the organized body of knowledge that is called a science seems to require measurement techniques for its development. The history of most areas of knowledge show that in the early stages of its development it is acquired by qualitative methods, without resort to measurement. Such knowledge is usually lacking in precision and often hopelessly vague, but the kernel of truth that it contains opens the way to the development of progressively more precise knowledge.

Sometimes these early qualitative observations are of immense importance. For example, the observations of Freud formed the basis for the development of much of clinical psychology, although many years had to elapse before the development of measuring techniques

and experimental methods permitted the systematic testing of aspects of his theories. Qualitative observations seem to be essential for the development of any branch of science, at least in its early stages, but it is ultimately careful work involving measurement that builds a science of real value. At a rudimentary level, even qualitative concepts can show some organization, as they did in the case of Freud's; but the ultimate test of the validity of such concepts is whether they do or do not facilitate prediction. Tests of the accuracy and validity of prediction almost inevitably involve measurement.

The statement that measurement is central to the development of a science is justified more by history than by logic. Each field that has become a science has shown a dreary period of slow advance prior to the introduction of methods of measurement.

Perhaps it would be well to pause here and reflect briefly on the role that measurement has played in science. Consider, for example, the well-known studies of the Abbé Mendel. He started out with the observation known to most farmers that crosses of different types of the same plant produce a new generation in which the characteristics of the parent plants may be distributed in various ways. Mendel was able to count the frequency with which each of the characteristics appeared in the offspring, and on the basis of these counts he was able to lay the foundation for a science of genetics. It is quite inconceivable that a mechanism underlying the inheritance of attributes would ever have been discovered without the use of such measurements. The crucial fact was one that involved quantity, namely that approximately 75 per cent of the offspring of a cross between dwarf peas and tall peas were tall.

Numerous other examples could be given of the dramatic role played by measurement in the founding of other areas of science. A science of mechanics came into being when Galileo was first able to introduce a simple way of measuring the rate at which bodies fell. Much of the work that Newton had begun came to a standstill for nearly one hundred years until that great experimentalist Cavendish was able to measure the gravitational constant. Lavoisier's careful measurement of the weight of the products of combustion demonstrated that the phenomenon of burning involved a combination of a substance with a component of the air, and this in turn revolutionized chemistry. Later, the measurement of atomic weights supplied the basis for laws of the combination of the elements. Almost every

major advance in science has been closely allied to the development of new methods of measuring or handling quantities.

Much the same seems to have been true in the behavioral sciences, of which research in education constitutes a part. Binet's attempts to measure intellectual power advanced immensely the possibilities of making predictions of behavior in ways that the earlier qualitative methods had been unable to do. When J. Maynard Rice first developed methods of measuring certain outcomes of the educational process, he made it possible to compare systematically one classroom procedure with another. E.L. Thorndike made it possible to exercise control over certain aspects of the curriculum by providing measures of the relative difficulty of various words, which in turn made it possible at a later stage to measure the difficulty of reading materials. Further development of evaluation techniques has made it possible to conduct research in education that simply would not have been possible fifty years ago. Even though the measurement techniques that have been introduced in education are crude, they have permitted a great expansion in our knowledge of the educational process.

Measurement involves the assignment of numbers to events according to some rule. The scale used at the post office for weighing packages assigns to a package a number that indicates its weight. The scale has been built to assign numbers to packages according to a rule prescribed by the National Bureau of Standards. A much simpler type of measurement is illustrated by the assignment of numbers to baseball players in order to label them. In the technical meaning of the term, the latter also involves measurement because it involves the assignment of numbers to objects or events according to a rule. Measurement, as the term is used in their book, is hence rather broadly defined.

Levels of Description

It used to be said not so many years ago that only insofar as the observer was describing specific acts was he describing behavior with any precision. Actually this is not true, for behavior can be described in a whole range of terms, from those that refer to the minutest detail to those that refer to gross total units of action. This concept is better understood by illustration. It is possible to describe behavior in terms of large units, such as, "The teacher showed the class how to carry out long division by working several simple examples on

the blackboard." It is also possible to describe the same situation in terms of smaller units of behavior, such as, "The teacher (1) entered the class, (2) said, 'Good morning, children. Today we are going to learn long division,' and (3) explained the general idea of long division and why it was useful to learn it, etc." Still further details could be given by describing the movements of each part of the teacher's body in space and in time, and by providing a record of the sound vibrations produced by her larynx. Such a detailed level of description of behavior probably could not be used because of the vast quantity of data it provides and the immense difficulty involved in handling such massive quantities of facts. On the other hand, the broad descriptions provided by observation of the gross over-all type provide too little data for the purposes of most research. All levels of description refer to behavior, and both stimuli and responses can be described in terms of a great range of complexity. What the scientist has to do is to choose the level of description that will ultimately permit him to make useful predictions. This may be said in a different way, which is explained elsewhere, namely, that the scientist must choose a data language suited to his purpose.

In this connection it is common to distinguish between "molar" and "molecular" approaches to research in education. This distinction comes from the field of chemistry, where the term "molar" refers to a rather large mass of material in contrast to "molecular," which refers to a small particle. When research in the behavioral sciences is said to be undertaken at a molar level, it means that it is concerned with gross aspects of behavior rather than with minute details. Research can involve the minute aspects of behavior, but this is not generally considered to be profitable. Hence, most educational research is molar.*

* When Hull originally considered his molar theory of behavior, he was inclined to think that all of the response dimensions such as speed of response, forcefulness of response, and others would all be related, and indeed positively correlated, dimensions. Empirical research has shown that this is not the case. It seems far more correct to assume that measures of two response dimensions will be correlated only when they are learned (reinforced) together. The Hull theory of the interrelationship of response dimensions, which has become known as the micromolar theory, avoided the need of measuring any other than the gross consequences of behavior. Considerable question has been raised in recent years concerning the soundness of the assumptions on which it is based, and some have advocated a micromolar approach that admits the necessity of the study of behavior in greater detail and that recognizes each aspect of a gross response as a separate response in itself, to be studied in its

The choice of the proper level of description is important in all aspects of educational research. If one is concerned with school plant and facilities, it is probable that one will not be concerned with the size of the brick used, but the classroom may be a convenient unit with which to deal. Then, again, all of the buildings in a school system would probably compose much too large a unit with which to be concerned. Much the same is true of behavior. A unit of behavior must be selected that is not too large nor too small for the particular purposes that the researcher has in mind.

If a study is being conducted in which pupils are required to read as quickly as possible words which are flashed on a screen, it is likely that the experiments will be concerned only with whether the words are or are not correctly read, and not with all the variations that may occur in how the word is said. The response involved in saying a particular word is an involved constellation of events, and the word may be said softly or loudly almost immediately on presentation of the stimulus word or after considerable delay; it may be spoken clearly or slurred; it may be said in a monotone voice or it may be said with fluctuations in pitch; and so forth. In addition, the reading of the word on the screen may involve a variety of processes. Sometimes the word may be recognized immediately, and sometimes its pronunciation is elucidated by means of the application of phonetic principles. In the molar approach we are not concerned with the myriad of variations that may be related to the making of a particular response. It is evident that when this is done, a certain amount of information is lost.

CLASSES OF VARIABLES IN EDUCATIONAL RESEARCH

In order to test hypotheses in ways that determine whether they should be rejected or accepted, it is almost always necessary to use

---

own right. This leads into a field in which relatively little is known. Little advice can be given to the student concerning the degree to which he should plan his studies on a macromolar or a micromolar level. Nevertheless it may be pointed out that, as research proceeds to a more detailed level of behavior, there is increasing difficulty in measuring the characteristics that it is desired to measure, and more and more complex laboratory instruments are needed. There are strictly experimental advantages in the study of molar behavior rather than molecular, and these exist quite apart from any theoretical advantages. From the point of view of the educator, it is molar behavior rather than molecular behavior that is of interest.

concepts that permit measurement. It has been found convenient to classify variables on three rather distinct bases, with which the reader should be familiar.

First, it is common to classify variables into the categories of dependent and independent variables. In experimental studies the condition that is varied is referred to as the *independent variable*. If the amount of time devoted to drill in spelling is varied in a study, then this is the *independent variable*. If the effect of drill is measured by means of a spelling test, then the score on the spelling test is referred to as the *dependent variable*. These terms have second meanings, commonly used by statisticians: the variable that is being predicted is called the *dependent variable*, while the variable from which predictions are made is called the *independent variable*.

Second, one may classify variables in terms of the phenomena to which they relate. Thus it is common to distinguish between variables related to the stimuli that impinge on the individual and those that are related to his responses.

Third, one may classify and consider variables from the point of view of their mathematical properties. While a very large literature has been written on this last type of classification, it will be touched on only briefly in this chapter because up to the present it has had only limited consequences for research methodology.

Each one of these three classifications of variables must now be considered.

DEPENDENT AND INDEPENDENT VARIABLES

Consideration must first be given to the primary meaning of the distinction between these two classes of variables, which derives from experimental psychology. The experimenter, whether in education or elsewhere, varies certain conditions in order to determine how variations in these conditions produce certain consequences. In most educational experiments, the experimenter varies some condition in the environment of the child, such as some aspect of the teacher's behavior. He then seeks to determine how this affects achievement as measured by a test or by some other device. The variation in the teacher's behavior is the independent variable, while the achievement score constitutes the dependent variable.

The second and broader meaning of these terms seems to have been derived from statistics rather than from any other source. In most scientific research, events are considered to occur in a time continuum, and certain events precede and are considered to be necessary antecedents of other events. The researcher usually measures certain characteristics of a situation as it exists at a particular time and tries to relate his findings in this respect to measures of previously existing conditions. It has become customary to refer to the variable to be predicted as the dependent variable, since is viewed as being dependent on previously existing conditions.

The type of event that it is sought to predict is occasionally an all-or-none affair, such as whether a person will or will not graduate from high school, or will or will not commit a crime. More often it is desired to predict some aspect of behavior that can be given a position on a continuum—as, for example, what grade a pupil will achieve in a particular course, or the liking a student will express for specific curricular materials, or the change in the absentee rate when a new building is provided. In such cases, and in most cases in educational research, the research worker desires to predict the value that a variable will assume under given conditions.

The variables that educational research is ultimately concerned with predicting are response variables. That is to say, they are characteristics of the way in which the person responds to his environment. This may not be apparent to many educational researchers, so further comment is necessary. Much curriculum research, for example, appears to have as its main goal the development of a curriculum. The curriculum developed exists only because the researcher believes it will have a better effect on the child's behavior than other existing curricula. In a similar way, research on problems of school plant are justified only because it is believed that the type of plant provided affects the behavior of pupils. Much research is based on the assumption that certain educational events affect behavior, and ultimately such assumptions must be tested even though this cannot be done at the present time.

The variables that are used for making predictions are commonly referred to as the independent variables of research. It is not the nature of the variables that makes them dependent or independent but the way in which they are used. Indeed it frequently happens that

the dependent variables in one study becomes the independent variables of another study.

In a study of predicting reading skill resulting from training in a foreign language, a measure of skill at the end of the period of training would probably be the dependent variable. The independent variables would be chosen according to the nature of the inquiry, and they might include various aspects of the way in which teaching could be varied and aptitude tests given prior to training. The independent variables might also include earlier conditions related to the ability to learn the particular language, such as exposure to related languages in childhood. They might also refer to such matters as the rewards, incentives, or reinforcements provided by the learning situation, or to any other condition related to the learning process.

The dependent variables of educational research, which ultimately are response variables, may refer to the frequency with which certain types of behavior occur or to qualities of behavior in particular situations. Often they represent verbal behavior, which, in a civilized society, is one of the more important aspects of behavior. The ways in which response variables are derived from the vast number of events that constitute the flow of behavior are discussed elsewhere in this book. It is sufficient to say at this point that many unsolved problems are involved.

The ultimate purpose of a science is to permit both the prediction and the control of events. Some sciences must be content with only the prediction of events, as is the case of astronomy, where control appears to present insuperable difficulties. The educational researcher is relatively fortunate in this respect, because not only can he expect to be able to predict events—such as who will succeed in college—but also he can aspire to exercise control over events—as when he seeks to develop a curriculum that will achieve a particular goal with maximum efficiency. At the present time, it is probable that the educational researcher knows much more about predicting events than he does about controlling the educational process to achieve particular ends. Prediction alone fulfills an extremely important function in education, and the capacity provided by research to do this has had a very important influence on educational practice. If it is possible to predict who will succeed in medical school and who will not, it may

then be possible to prevent many from experiencing the frustrations of failure. If it is possible to predict who will become delinquent unless remedial action is taken, there is a possibility of making a radical change in certain aspects of our culture. If one could predict which students of education will become highly neurotic, if not psychotic, teachers, a major problem of educational administration would be solved.

While prediction performs an extremely important function in education, an even greater contribution could be made by educational research if it could tell the educator how to arrange conditions in order to produce specific results. Some progress has been achieved in this direction. A little is known about the types of classroom conditions that are effective for particular purposes. Some knowledge is available about the effects of particular types of classroom organization. But knowledge concerning the control of educational conditions is still fragmentary because it is hard to acquire. The reader will note in subsequent chapters that it is generally much easier to conduct studies that lead to prediction than to conduct those that lead to control.

Here again the reader should be warned that the conditions which produce effective learning in one pupil will not necessarily produce effective learning in another. When it is said that the function of the educational research worker is to discover the conditions related to effective learning, it is assumed that these will be such that they can be modified as required by individual differences. The fact that children come to the classroom from different backgrounds and with different abilities results in their responding to the learning situation in different ways. Our knowledge of conditions related to effective learning should make us flexible in prescribing for particular pupils; it does not imply any rigidity of educational practice. In addition, teachers vary in the conditions they themselves are capable of producing. For example, some who may be otherwise capable are unable to generate a warm, friendly atmosphere. Such teachers will probably have to use somewhat different teaching techniques from those of teachers who have sympathetic dispositions. All such differences among persons, whether pupils, teachers, or others, must be included among the variables with which educational research is concerned.

CLASSIFICATION OF VARIABLES IN TERMS OF THE PHENOMENA TO WHICH THEY RELATE

I. Stimulus Variables

An important classification of variables that has had great influence on the language of current theories of behavior is that in terms of stimulus variables, response variables, and intervening variables. Ten years ago this classification could have been glossed over in a few lines and summarily dismissed, but its current importance is such that it cannot be treated so briefly today. Over the years it has become necessary to specify with increasing precision what is meant by these classes, and although full agreement has not been reached, it is important that tentative definitions be given here.

The first class of variables that we must consider here is the stimulus variable. The term *stimulus*, or the corresponding quantitative term *stimulus variable*, we now recognize does not define a clear-cut class of phenomena, as it was once thought to do. The best analysis that the writer has yet come across concerning the varied senses of the term is supplied by Verplanck (in Estes *et al.* 1954), who distinguishes four usages as follows:

Usage I. In this usage, the term refers to a part of the environment or to a change in the environment. In this meaning, a stimulus is considered to be a stimulus even though it has no observable effect on the organism. Since most psychologists are interested mainly in those aspects of the environment that affect behavior, this meaning of the term includes a wider range of phenomena than is usually needed.

Usage II. This is the usage most commonly found in textbooks of physiology and in most general textbooks on psychology. It refers to any form of energy that produces a response in the receptors. What constitutes a response is often a moot question, for it is also common for psychologists to refer to subthreshold stimuli, which individually may not produce a response but which collectively do. Thus a light of low intensity may not produce a retinal response when it is of one millesecond's duration, but five similar successive one-millesecond flashes may do so. The distinction between forms of energy that produce a response and forms that do not is far from being as clear cut as it is presented to be in elementary texts.

Usage III. This is the usage that is implicit in most educational research. In this sense, which is also accepted by B.F. Skinner in studies of animal learning, a stimulus is not considered as such until the subject manifests an observable behavioral response to it. Thus a black square on a white background is not regarded as a stimulus in many rat studies until the animal has learned to respond to it in some way, as by moving a lever or pushing the square back to enter a food box. In a similar way, one would not usually refer to a chart on a wall as a stimulus if the children never showed any evidence of responding to it. A stimulus variable in this sense of the term would be described as *response-inferred*. This is an undesirable property in many ways, since the stimulus is inferred to be a stimulus through a rather remote chain of events. On this basis, many psychologists believe that a rigorous science of behavior would not include such remotely inferred classes of events.

Usage IV. In this usage, the stimulus is identified as in class II with a physical event, but the event occurs within the organism. Further, the physical event is not identifiable as it is in class II, and is inferred only from behavior. Until the event is actually identified, it must be considered to be hypothetical—that is to say, an unobservable event postulated to account for observed behavior.

In listing these different ways in which the term stimulus is used, Verplanck states that it is most unfortunate that a single word should be used to refer to all of these phenomena. It should be noted that only in the case of the first two of these usages does the term refer to a class of environmental variables that the experimenter can vary. The researcher should be aware of this fact, and when he himself refers to a stimulus or a stimulus variable, he should know which one of the various usages he has in mind. He should not perpetuate the confusion that has existed in the use of the term "stimulus."

The definition that we prefer to accept at this time is the first one given by Verplanck, except that in educational research we are concerned with only those environmental conditions that may be *hypothesized* to effect behavior.

Examples of Stimulus Variables in Educational Research

Any measurable aspect of the pupil's environment that in some way may be expected to affect his behavior is considered here to be a stimulus variable. These aspects include not only the characteristics

of the buildings, textbooks, visual aids, and other features of the physical environment, but also the behavior of teachers, counselors, parents, principals, and others with whom the person seeking education comes into contact. So far, relatively little effort has been devoted to the isolation and measurement of stimulus variables related to the educational environment, but a few may be mentioned by way of illustration. A first example is presented by measures of reading difficulty such as those provided by the Lorge index or the Flesch index. These measure an important property of printed verbal material. A second example of a significant stimulus variable with respect to the pupil is the amount of aggressive or dominant behavior manifested by the teacher. Studies at Michigan State College have shown clearly how this variable is related to the behavior of pupils.

Relatively little is known, however, about the relevance of physical aspects of the surroundings of the child, and most relevant stimulus dimensions that have been defined refer to the behavior of persons with whom he comes into contact. It is hardly surprising that this is so, since the major influences in a person's life result from contacts with other persons rather than from contact with physical events as such. This may be due to a very great extent to the fact that educators have not yet learned how to arrange physical conditions to maximum advantage. Recent studies in the teaching of foreign languages have shown that much is to be gained by isolating students in booths where they are instructed by tape recorders. In this learning situation, not only is the student isolated from distracting stimuli, but also it is possible for him to repeat parts of the lesson on the tape recorder as many times as he wishes. While research in the past has emphasized the importance of the role of the teacher and the resulting interpersonal relations, educational research may become more and more preoccupied with the physical conditions related to learning.

Many studies of classroom learning are concerned with a whole range of stimulus variables represented by conditions other than teacher behavior. Of particular interest at the present time are so-called visual and auditory aids, which are extensively used although there is not much evidence to justify their use.

The growth of the use of visual and auditory aids unfortunately has been influenced by factors other than those derived from rational

planning. Since visual aids are an easily observable but probably superficial aspect of the classroom environment, there is always a temptation for the teacher to cover the walls of the room with charts and other devices, for these may be the basis on which parents, school board members, and even supervisors may judge the merit of the teacher's performance. While this basis for judging a classroom is seen frequently enough in civilian schools, it is probably the most common way of judging teaching effectiveness in military establishments, where the inspection system is such that major emphasis is placed on superficial detail. This emphasis on the importance of what can be easily observed has stimulated not only the worst features of the current use of visual aids but also some of the best features.

The most extensive research that has been done in this area is in the use of moving pictures. Investigations of this sort go back to some studies in Europe in the early 1920's and substantial work done by L.L. Thurstone and his associates in the 1930's. There has also been considerable work done since that time as a result of the widespread use of films for educational purposes in the armed services during World War II. This field of research is a promising and developing one, but those who enter it should be forewarned of some of the difficulties.

First, a problem is presented by the fact that most visual aids are developed for the purpose of achieving some specific goal that could be achieved as easily by other means. If one does not know just what is to be accomplished by a particular visual aid, it becomes impossible to determine whether it is or is not effective. Most moving pictures fall into this category. For example, during World War II, a series of films was developed entitled *Why We Fight*. Considerable efforts were made by psychologists to evaluate these films as devices for producing attitude change, but it was soon found that the films had not been made with clear purposes in mind, and that, as a result, they seemed to contribute to the development of a whole series of rather independent attitudes.

Second, it is suspected that most visual aids do not present a sufficiently extended experience to produce either measurable changes or changes of any great permanence. The writer's prejudice is that unless a whole program of visual aids is directed towards a specific

goal that little is likely to be achieved and that no measurable changes in level of achievement will occur.

Third, most visual aids are not related to the learning process in a theoretically sound way. Often they are only knickknacks designed to enliven the classroom in some manner, as if such enlivenment would necessarily have any effect on the course of learning. The same mistaken notion is that an irrelevant joke somehow makes the circumstances of learning more congenial. A joke in the classroom is often just a substitute for effective teacher behavior. In the case of many visual aids, one suspects that the student remembers the gimmick but not the point that the gimmick is designed to illustrate. One may remember a cute gadget while forgetting the meaning it was supposed to convey. This is a matter that could well be investigated.

Fourth, it is suggested that research on visual aids should attempt to evaluate not the effects of specific devices, but those of groups or programs of visual devices that collectively are designed to achieve a particular goal. Under such conditions, it is possible that the effect of the devices may be sufficiently extensive to produce measurable effects.

Visual and auditory aids have been considered as manipulable aspects of the pupil's environment. In this sense, the total curriculum may be considered to be an extended program of this type, since the visual and auditory senses are the main avenues through which curricular materials have an impact on behavior. In view of what has been said, the best possibility of obtaining differences in pupil behavior as a result of differences in educational treatment would be when two matched groups of students are exposed to fundamentally different curricula. However, there are difficulties in this approach, which need brief consideration.

First, there is the difficulty of locating curricula that differ in any fundamental and permeating respect. Social pressures are such that most curricula contain a common core of material required by society if not by school boards. Few bodies that control education permit much deviation from the social norm in the design of curricula.

Second, it is often difficult to determine just how curricula do differ. We must not be fooled by names and assume that, just because they differ in name (as do the so-called traditional and progressive curricula), they differ in significant stimulus characteristics as far as

the pupil is concerned. The results of much research on curricula in the past have had little validity because the differences in curricula have been differences only in name. What is needed is measures of various properties of the curricula. This is feasible in many cases, for example when measurements are made of the amount of time spent in various activities such as lectures by the teacher, class recitations, group activity, individual study, and so forth. The main difficulty in obtaining such measures is that of observing classrooms without altering the process that is observed. One suspects that, if observation is sufficiently prolonged and is extended over a period of more than a week, the teacher and class will become accustomed to the presence of an observer and will show a typical pattern of behavior.

The difficulty of quantifying differences between curricula stems partly from the fact that quantification must depend partly, if not largely, on judgment. Detailed analysis of recordings of what happens in the classroom is not a feasible venture because of the vast amount of material that it entails. Although it is possible to obtain rough quantitative estimates of the time spent by the teacher in lecturing, in conducting recitations, etc., it is not easy to quantify by present methods the time spent by the teacher in imparting information, in asking questions, and in correcting pupil's products; hence these must usually be estimated by the observer and the results of the estimations recorded on a rating sheet.

In summary, while a science of education requires that measures of important characteristics of the environment believed to have relevance to the educational process should be available, there are as yet few aspects that can be measured satisfactorily. At least a part of this difficulty stems from the fact that most other branches of the behavioral sciences have devoted little of their effort toward the measurement of environmental conditions related to behavior.

II. Response Variables

The ultimate purposes of education are defined in terms of desirable ways of responding to life situations. It is clearly not enough for the educator just to believe that he has produced certain internal changes in those who have passed through his educational program. It is generally conceded that the success of any educational program lies

in its effect on behavior in those situations that the program has been designed to help the pupil face. If a program has among its objectives that of developing a critical attitude toward political propaganda, it is a failure unless in later years the person who has passed through that program shows through his behavior a critical attitude toward political propaganda. It is always through responses that the success or failure of an educational program can be established.

The common approach to the measurement of the way in which a person responds to his environment is best understood if we turn to some of the classroom responses that are commonly accepted as evidence of achievement. A useful illustration is provided by the teaching of reading, in which it is common in the early stages to build up a recognition of fifty to one hundred everyday words. The pupil becomes able to recognize these words although he has no knowledge of phonetics, and this recognition enables him, with a little help, to read simple books designed around them. The pupil's recognition ability for these words can be measured in terms of the number he recognizes and speaks correctly. His learning in this respect is measured in terms of the frequency with which he makes a correct response in a standardized situation. Sheer frequency of response is one of the commonest types of response variables that the educator is likely to encounter. Many personality variables are measured in terms of frequency of occurrence of a particular response. Sheer frequency of occurrence of a particular class of behavior is used as a major method of measuring response characteristics, not only in education but throughout psychology. Not only may we measure the progress of learning in the case of a rat that is acquiring the skill of running a maze by determining the frequency of occurrence of errors (wrong turnings), but similar techniques may be applied to charting the course of learning of the pupil. Aspects of the pupil's personality development may be measured by counting the number of antisocial acts shown toward other pupils, or the number of acts of hostility shown toward the teacher. Frequency of occurrence of many kinds of maladjustments is often used by counselors as an important item of information to be used in helping the pupil.

The behavioral variable of educational research may be the frequency with which a particular event occurs, or it may be the scale value of a particular event. Frequency measures of identical events.

such as the frequency of saying "No!" or the frequency of physical aggression, are easily understood. More difficult to understand are measures derived from scaled systems of events. The scale value of a particular event is a more complex concept. It may be illustrated by studies of racial attitudes, in which persons are required to respond to some situation involving a member of another racial group. The response may vary from positive and friendly to negative and hostile. The response may be assigned by one of a number of procedures to a position on a numerical scale that varies between the extremes. Another example of a scaled response is the response to a vocabulary test as measured by the total score. Such a test presents a series of words of graded difficulty, and the person taking it may be expected to define correctly all of the words up to a certain point but none beyond that point; thus the point where passing an item changes to failing an item can represent the number answered correctly and also the position on the scale that represents the maximum level of difficulty of the words that are successfully defined. Here again various methods may be used for assigning scale values. However, it should be noted that the vocabulary test described here represents an idealized situation that is unlikely to be duplicated in actual practice. What happens in a well-constructed vocabulary test is that, although there is not a completely sharp break between the point of passing and the point of failing all items, there is a limited zone within which this break occurs in a rather irregular fashion. Scaling is approximate rather than precise in such a situation. Behavior theory development, insofar as it has been attempted on a rigorous basis, has been concerned with response-frequency measures or their correlative response-probability measures.

In research in education there are many frequency-of-response variables of great significance. In the early stages of reading, the frequency with which particular words are recognized represents an important class of variables. Frequency of errors in written compositions in English or in foreign languages is one of the commonest variables measured by teachers in those areas. They are not concerned with the level of seriousness of these errors, but with the number of times they occur within a given range of opportunity. Computational errors are also of this type. The arithmetical operation 9×9 is performed perhaps one hundred times by a child, and on 95 per cent

of the occasions he performs the operation correctly, but from time to time, even though the response is highly overlearned, an error occurs. One might say that the probability of a correct response in such a case is 0.95. In other children the correct response probability is perhaps 0.75 or 0.42. There are differences at any given time in the response probability for groups with exposure to equal amounts of training. In practice, we are likely to be interested in predicting a sum of such response probabilities, as when we administer a test of one hundred simple computational problems from the multiplication tables and count the number of errors that are made. In measuring computational skill, this procedure is more likely to be adopted than is the procedure involved in preparing a set of problems of graded difficulty—which, in this case, would be problems graded in terms of complexity.

This discussion is presented to show that variables that represent examples of response frequency or response probability are commonly used in educational research. It should not be taken to indicate that variables in which a score represents a scale position should be avoided in the development of educational research. The latter may be difficult to fit into any current system of theoretical psychology.

In the case of response variables, great difficulties are also encountered in establishing scales in which the units can be considered to be equal in any way. Attempts have been made to scale responses in various ways so that the resulting scale can be considered to consist of equal units, but objections can be raised to all of these systems. Many psychologists would accept the viewpoint expressed by Alfred Binet that in the measurement of responses all one can do is to rank individuals and that true scaling is rarely if ever possible in this domain. This problem will be discussed at greater length later in this chapter.

III. Intervening Variables

A few psychologists have taken the view that a science of behavior could be built simply by studying the relationship of stimulus variables to response variables. Such psychologists assume that response variables are direct functions of stimulus variables. If this were the case, education would simply be a matter of arranging and scheduling stimuli so that the desired responses were elicited. On this basis, a

theory that accounted for behavior in the early stages of learning to read might perhaps consist of two postulates, which might be worded as follows:

Postulate I: The probability that a correct response will be made in recognizing a word is directly related to the number of times the correct response has already been made.

Postulate II: The probability that a correct response will be made on a specific occasion in recognizing a word is inversely proportional to the time elapsing between practicing the response and measuring the probability of its occurrence.

This miniature theory, which covers the early stages of learning in which facility is acquired in the recognition of common words, is really a practice-makes-perfect type of theory. It states in a straightforward manner that the ability to recognize words is simply a function of practice. Every schoolteacher in the first and second grade can testify to the inadequacy of such a theory and can show why it is wrong. Pupil after pupil in the first grade shows no improvement at all in recognizing words, even though extensive practice is given. The simple deduction from the theory that the child who has to build up a recognition vocabulary should simply be given more practice, which will inevitably remedy the deficiency, is a deduction that is just not in accordance with the facts.* Some pupils in the first grade are not capable of reading, and hence a theory of reading that has at least minimum adequacy must include the concept of capability, an internal condition that is not directly observable and that accounts for individual differences in the responses of children to equal amounts of word-recognition practice.

Even in the simplest cases, it is not possible to describe events merely in terms of the relationship between stimulating conditions and responses. Consider, for example, the case of the eyelid reflex produced by a slight puff of air on the cornea. If the puff is very light,

* If the theory were to be used for any other purpose than to illustrate a few points in the immediate discussion, it would be necessary to define the terms used in some detail. Terms such as *response probability* and *practicing a response* would need careful definition if the theory were to be used as a basis of research. The theory is used here only to illustrate the absurdity of one that uses only response variables and stimulus variables.

the response may or may not occur, depending on the condition of the individual. Responsiveness varies according to internal conditions such as fatigue, attentiveness, the degree to which certain chemicals are present, and so forth. In the case of more complex behavior, the intervening conditions cannot be identified with any known chemical condition or identifiable neurological structure. The intervening conditions that must be postulated therefore are said to be hypothetical. Whatever varies when these intervening conditions are varied is referred to as an *intervening variable.**

Intervening variables have been described as hidden variables. In more technical terms, they can be said to have transcendental properties, in that they represent conditions that transcend experience and cannot be observed directly. Aptitude for reading or reading readiness is not a property that can be observed with any directness in a child. All that can be observed are the results of this variable as they are manifested in the task of learning to read.

The reader should be warned against the introduction of intervening variables by circular argument. The fact that children differ in the rates at which they learn to read is not a sufficient basis for inferring a variable referred to as reading aptitude. It helps little to postulate the existence of an intervening variable and then to use this variable to explain the differences on the basis of which it was postulated. Such circularity serves little purpose. On the other hand, if it can be shown that equal conditions of learning still result in individual differences in reading skill, the investigator is on rather firmer ground in postulating such a variable. Indeed, some variable must be postulated to account for the lack of relationship between the stimulus variable and the response variable. The theory would be on still firmer ground if it could be established that a variable measured by a certain specific test (which was not a reading test) could be used as if it were a

* The terms *intervening variable* and *hypothetical construct* (or *construct*) are used by many writers almost interchangeably. At one time an attempt was made by Meehl and MacCorquodale (1948) to draw a clear distinction between these two concepts, but objections were raised to the proposed distinction that they made. The suggested distinction was that a *hypothetical construct* involved ideas over and above those included in an *intervening variable*. Much of what is referred to as a hypothetical construct in current psychological theory represents mechanisms that are postulated to mediate between stimulus and response. These mechanisms are not observable and should really be considered as imaginary mechanisms that it is convenient to postulate. The utility of such devices for guiding the thoughts of the scientist will depend on the extent to which they lead to variables that have predictive value.

measure of this aptitude. If, for example, a measure of physiological maturity were to account for individual differences in reading skill after exposure to equal training conditions, it might be said that this variable operated as the intervening variable that must be introduced to account for reading behavior.

Intervening variables may refer to a wide range of conditions. Sometimes they may have a clear relationship to tissue functions, as when a condition of hunger is produced by food deprivations over a period of hours. Studies of changes in behavior in pupils at various stages of food deprivation have been conducted, in which it has been shown that behavior changes as food deprivation is increased. The condition of deprivation may be considered to be a measure of an intervening variable in such experiments. An experimenter might, for convenience, refer to the condition of food deprivation as a condition of hunger, but by doing this he is likely to confuse the issue. While children deprived of food for a period of four hours might refer to themselves as hungry, it is quite possible that the vague inner state of discomfort of which they were aware might be of only minor importance in modifying behavior in comparison with the effects of deprivation of which they were not aware. For this reason, the wise experimenter would do well to refer to this variable as deprivation rather than hunger. He would also do well to measure deprivation by the number of hours without food rather than by any subjective estimation of hunger.

By the example given, it is not suggested that research on food deprivation in children is likely to be particularly profitable. It is merely a convenient example of an intervening variable that illustrates many of the problems of defining such variables.

A major class of intervening variables is the so-called aptitude variables. These are presumed to measure conditions that facilitate learning, and there is substantial evidence that many of them measure facilitations produced by previous learning. It is possible that some measure rather permanent conditions in the nervous system, which have been influenced only to a small degree by learning. Relatively little is known at this time about the generation of these variables, but the fact that they can be measured and that these measures have had predictive value have given them a position of the greatest importance in educational research. One of the rather surprising facts about

this class of variables is that relatively few aptitudes that have definite predictive value in learning situations have been discovered.

Motives also represent intervening variables, and they also represent variables that are commonly but erroneously postulated on the basis of direct observation of behavior. Strictly speaking, motives are unobservable. If two persons perform the same task but at different rates, there is no value in stating that the one is better motivated than the other, for by this is meant only that the one worked more rapidly than the other. A beginning has been made in the measurement of motives by independent techniques, and also a beginning has been made in a useful classification of human motives. At one time it was thought that interests might constitute powerful measures of human motivation, but these promises have not been fulfilled.

Many variables that we have considered here as stimulus variables may be treated as intervening variables. For example, we have treated variations in training as stimulus variables in this chapter. We could have viewed them as external conditions that generate internal conditions, which in turn operate as intervening variables in an actual reading situation. When the child is tested for his ability to recognize words, the printed words are the stimuli, what he says when they are presented constitutes the responses, and internal conditions—some generated by previous training and others by complex conditions including heredity—represent the factors that produce intervening variables. The latter is just a different way of viewing the situation, and it may be just as satisfactory a way as that discussed.

CLASSIFICATION OF VARIABLES IN TERMS OF THEIR MATHEMATICAL PROPERTIES

Students of psychometrics have emphasized the importance of the mathematical properties of psychological measuring devices, because these properties determine the operations that can and cannot be performed legitimately with measures derived from them. The classification commonly described is that given by S.S. Stevens (1946), and it will be briefly outlined here. Modifications of this system by Coombs (1953) add the concept of a partially ordered scale, which will be briefly mentioned.

At the least powerful level of measurement is the *nominal scale*,

which is simply a system of assigning number symbols to events in order to label them. The usual example of this is the assignment of numbers to baseball players in order to identify them. If these players were arranged in order of the numbers on their shirts, the order would have no meaning. Thus the numbers cannot be considered to be associated with an ordered scale, for their order is of no consequence. In educational research, it is quite common to identify events by numbers. Thus in observing a teacher, all hostile gestures of the teacher may be recorded by placing a check mark against the number 11, while words of praise administered by the teacher are indicated by checks against the number 15. The numbers are just convenient labels for the particular class of events. In these scales, the numbers refer to events that cannot meaningfully be placed in some order.

The lowest level of the *ordered scale* that is commonly used or discussed is the ordinal scale.* The *ordinal scale* places events in order, but there is no attempt to make the intervals of the scale equal in terms of some rule. *Rank orders* represent ordinal scales and are the commonest used in educational research.

Stevens distinguishes between two types of scales in which the intervals can be said to be equal in some way. These are *interval scales* and *ratio scales.***

In the case of the interval scale, the intervals are adjusted in terms of some rule that has been established as a basis for making the units equal. The units are equal only insofar as one accepts the assumptions on which the rule is based. For example, in psychophysics, it is common to accept as a unit the smallest difference that can be perceived. This unit is referred to as the differential threshold. Interval scales may have an arbitrary zero, but it is not possible to determine for them what may be called an absolute zero.

But ratio scales do have an absolute zero of measurement. The term "absolute zero" is not as precise as it was once believed to be. We can conceive of an absolute zero of length, and similarly we can conceive of an absolute zero of velocity. One object would have an absolute zero of velocity with respect to another when the distance

* Coombs introduces an intermediate category, *partially ordered scales*, in order to cover certain psychometric measuring devices that fall between nominal scales and ordered scales.

** Coombs also distinguishes *ordered metric scales*, which might be considered as a class between ordinal scales and interval scales.

between the two objects remained constant. Nevertheless, an absolute zero of temperature is theoretically unobtainable, and it remains a concept existing only in the scientist's mind. One could also change the present system used to number calendar years so that the zero on the scale would not approximate the birth of Jesus but would approximate the beginning of the universe (based perhaps on an expanding monoblock theory). Many assumptions would have to be made in establishing such a system for numbering years, and such assumptions would be more tenuous than those involved in the establishment of a zero of temperature. By means of even less acceptable assumptions, it would be possible to establish a scale of intelligence that would have an absolute zero *if the assumptions were accepted*, but there would be many who would balk at the assumptions. Indeed, so many would reject them that the enterprise would not be considered worth while. In passing, it is of interest to note that E.L. Thorndike suggested at one time the establishment of a scale of intelligence that would have something approximating an absolute zero.

This discussion serves to point out that an absolute zero, for the most part, is a concept rather than a reality. Whether a scale that claims to have an absolute zero is accepted as such depends on the general acceptability of this concept to those who might use it. The actual existence of such a zero is not always demonstrable, nor is it necessarily some directly observable condition.

While measurement involves the assignment of numbers to events, when these numbers can be considered to represent a scale, then it is clear that the events have some order. Thus the scientist who selects events that have these properties has thereby succeeded in perceiving some order in the general domain that he is studying. Measurement may thus in itself be an ordering process, and it is this kind of ordering that in the past has been particularly effective in enabling man to exercise some control over his environment.

At the highest levels of measurement, it is also possible to perform certain arithmetical operations with the measures, such as addition, subtraction, multiplication, and division. However, it is obvious that, in the case of scales that have no absolute zero, one measure cannot be multiplied by two in order to obtain a measure indicating a quantity twice as great.

One cannot justify performing any arithmetical operations with

most of the scales of measurement used by the psychologist. One cannot assume that a mental age of eight years on the Binet type of scale represents a level of intelligence twice as great as that represented by a mental age of four years. It is also unreasonable to say that a standard score of 60 on a standardized achievement test represents twice as much knowledge as a standard score of 30. It is just not possible to make such direct arithmetical comparisons between scores on most tests that are available today.

Some Problems of Scaling

Certain aspects of scaling that have not ordinarily been considered in the past have appreciable consequence for educational research. Attention was first drawn to this matter by Lloyd G. Humphreys (1956) in an address to the annual testing conference sponsored by the Educational Testing Service. He pointed out that Cattell many years previously had noted that scales could be divided into two categories, normative and ipsative.*

Normative scales are represented by intelligence tests and tests of achievement, and generally by tests in which a scale consists of a distinct set of items where the total score represents some function of the number of correct and incorrect answers. Scores on normative scales are usually interpreted with respect to the performance of persons collectively described as a norm group. On a set of normative scales such as are represented by an achievement test battery, a person may have all high scores or all low scores, and we are interested in considering each score separately and independently of all other scores.

In contrast, *ipsative scales* are illustrated by those derived from the *Study of Values.* This instrument attempts to provide measures of the extent to which each of six values influences a person's life. These six value systems are the economic, the religious, the social, the scientific, the political, and the aesthetic. The test is set up so that the person taking it must express a number of preferences for one or the other of these values. Where he is asked to make a choice between a decision based on religious values and a choice based on economic values, he must choose one or the other; he cannot choose both. The result

* Some years later C.H. Coombs drew a similar distinction between relative and irrelative measurement.

is that if he tends to choose religious values, scores on the other values tend to be depressed. The scores are derived in such a way that the average scores on all scales are the same for all persons. A person scoring high on some scales must be low on others. The scores can thus be used for ranking the values for a single person. They compare the strength of one value with another *within that person.* They do not permit comparisons of one individual with another. This fact has certain important consequences when these scores are correlated with other variables that it may be desired to predict, and this is a particularly significant point to notice when tests of interest are used in an attempt to predict performance in some activity. Since many tests of interest present ipsative scales, it is as well to remember that in predicting performance with such tests we are correlating ipsative scales with normative scales such as class grades or with other measures of achievement.

When ipsative scales are used, they should be used for predicting ipsative characteristics. A test designed to provide ipsative measures of interest in school subjects should be used for predicting an ipsative criterion, such as the rank order of the success achieved by pupils in these subjects. The rankings on the interest scales and the rankings on the achievement scales would be compared for each person included in the study. It is possible that the interest rankings might predict the achievement rankings for some persons and not for others.

Another aspect of scaling that has particular importance for educational research is the difference between altitude and breadth. The distinction is important in building achievement tests, and it can be explained fairly easily in terms of the problem of building an achievement test in American history to be given to college freshmen. Such a test might call for information concerning the major facts of American history from the days of the early settlers up to modern times. It would not call for any of the more obscure facts of history but would measure the breadth of a person's knowledge of the important facts.

A different test in American history could also be made covering the same period. The latter could be divided into sections, such as the colonial period, the revolutionary period, etc. Within each of these periods, questions could vary from those that call for commonly known facts to those that pertain to relatively obscure events. A per-

son answering the questions within any one of these sections would tend to answer questions up to a certain point and then fail items beyond this point. Such a test measures something different from breadth of information, and what it measures may perhaps be termed altitude.

Oddly enough, those who construct tests do not seem to be too concerned over whether they are measuring breadth or altitude. Usually no mention is made of this in test manuals, and it is necessary to examine the test in order to determine whether the emphasis is on the one or on the other. As a result, relatively little is known about the relationship of breadth scores to altitude scores.

Does the Research Worker Predict Behavior?

An ambiguity commonly occurs in referring to behavioral measurement. Current usage is to say that we measure behavior. What in actual fact we do measure is *some property of behavior*. When the researcher states that he is measuring the behavior of the teacher in the classroom, he is really measuring only certain properties of behavior. To state that the behavior of the teacher is being measured carries with it the implication that the measurement supplies a complete description of behavior. In actual fact, the researcher is likely to measure, and that only rather roughly, certain limited aspects of behavior. From the measures themselves, it is possible to reconstruct to only a very limited degree what happened at the time when the measurement was made. We are never likely to be able to perform the literal function of predicting teacher behavior in the classroom. Nevertheless, this should not discourage us, for what is important is to predict not teacher behavior as a whole but those aspects of it that have some crucial effect on pupil behavior.

Even if it were possible to predict accurately every aspect of teacher behavior, there would be no particular purpose in doing so. Much of what can be observed represents a great range and variety of events that have little bearing on what is accomplished. A teacher who is restless and moves around the classroom may be displaying merely the consequences of a long period of sedentary work. The activity may function only as a means of restoring the circulation in the muscles of the legs. It is clearly a phase of behavior that there appears to be no particular use in predicting. The same is true of pupil behavior.

We are not concerned with predicting the numerous isolated actions that the pupil may perform after he leaves school, for most of these are specific responses to incidental situations and have only immediate consequences. On the other hand, we are very much concerned with the prediction of trends in behavior and characteristics of behavior that appear in a wide range of situations.

An analogy that has been commonly used in discussing this problem is that of transmitting information over a circuit. In any circuit, a fraction of the energy transmitted is relevant to the transmission of the information; the remainder represents unorganized and irrelevant energies introduced into the circuit by unknown and uncontrolled sources. The latter unwanted source of energy is commonly referred to as the "noise." In a good circuit, the energy changes are mainly signal energy changes and the noise energy is reduced to a minimum. In the analogy under discussion, the signal energy is compared with the aspect of behavior that it is desired to measure, while the noise represents the stream of minor behavioral events that have no consequence insofar as the building of a science of behavior is concerned. As a matter of fact, this has been used as a basis for a general theory of behavior developed by Miller and his associates (1955). Their theory is much more comprehensive and, as it is said, global than any type of theory that the writer would recommend as a basis for research, but it is of interest to note the possibilities that communication theory has as an analogy to behavior. Also, it may be mentioned that workers in fairly specific research areas have used the communication-theory analogy and have built miniature theoretical systems on this basis.

Summary

1. The development of most sciences has usually been accompanied by the introduction of quantitative methods.

2. The scientist must decide the extent to which he is to study details or gross events. He chooses a level of detail that is convenient and appropriate. In educational research it is molar behavior that is likely to be examined.

3. It has become customary to refer to the variable that is predicted as the dependent variable and the variable that is manipulated or used for making predictions as the independent variable.

4. Variables that are predicted in educational research are usually responses to the environment—either the frequency with which certain responses occur or the characteristics of the responses.

5. A stimulus is defined for the purposes of this volume as a condition existing in the environment, which is hypothesized to produce a response in an individual. Educational research restricts its interest to those environmental conditions that are hypothesized to affect behavior. Studies in the field of curriculum are studies of environmental conditions and their relationship to the learning process. Studies of visual aids are studies of the manipulation of stimulus conditions.

6. Studies of visual aids present certain difficulties that are likely to interfere with positive findings. The main difficulty in such studies is that they are concerned with very limited aspects of the learning process.

7. Differences produced in learning by differences in curricula are more likely to produce positive findings than are studies of the effect of specific learning devices. However, curriculum studies are difficult to undertake, and differences between curricula are difficult to measure.

8. The central difficulty in the conduct of curriculum studies, apart from administrative difficulties, is the measurement of differences between curricula. Curricula may differ in rather complex ways, and their characteristics cannot usually be measured in a simple manner.

9. The child's and the adult's responses to his environment are commonly measured in terms of the frequency with which a particular response is made. This is generally the simplest type of measurement procedure.

10. A more sophisticated procedure for measuring responses is the use of an ordered scale. Achievement is commonly measured by tests that approximate such scales.

11. Intervening variables represent characteristics of the person that influence behavior. Most of these characteristics cannot be observed directly and hence have been referred to as hidden variables. Such variables are used to predict individual differences in performance to learning situations, and they include intellectual aptitudes as well as motivational characteristics.

12. Scales may also be classified in terms of their mathematical properties. These properties determine the extent to which various operations such as subtraction and addition can be performed meaningfully, with scores derived from them.

13. Normative scales are used for comparing the performance of one person with that of another. Ipsative scales compare a person's performance in one area with his performance in another area. Normative

scales should be used for predicting normative characteristics. It is not meaningful to predict normative characteristics with ipsative scales.

Some Problems for the Student

1. A well-known psychiatrist developed a test that, he claimed, could be used for identifying certain types of mental abnormality. A research worker, interested in the problem, administered the test to groups of persons who were characterized by these abnormalities and in whom these abnormalities had been properly diagnosed, and also to a "normal" control group. He interpreted the test results according to the rules provided by the psychiatrist but found no relationship between the results and the presence or absence of particular abnormalities. However, when the test results were given to the psychiatrist alone and without further information, he was able to make a very accurate prediction of which patient suffered from which abnormality. What hypotheses can you suggest to account for this situation?

2. List some of the characteristics of textbooks that result in the facilitation of learning. How might these characteristics be measured? Can they be measured without devoting an excessive amount of labor to the task? How could the reliability of these measured characteristics be estimated?

3. A teacher suspected that many of the academic difficulties encountered by the children in her class could be attributed to the unfavorable home conditions under which homework was accomplished. In order to study this problem, she administered a questionnaire to the children, asking about the extent to which there were distractions such as television and radio going on while they were doing their homework. A score derived from this questionnaire was then correlated with school grades. What is wrong with this procedure for studying the problem?

6 The Use of Multiple Observations in Measurement

The Single Observation

Most inferences that can be made as a result of educational research are based on multiple observations. From observing a child respond to a single problem assigned by the teacher, no responsible person is likely to make inferences concerning the child's educational achievement level or his scholastic aptitude. The single observation provides insufficient information for making either one of these inferences. Multiple observations can be used for making fairly accurate inferences about a child's achievement. Achievement tests are such means of providing multiple observations about a child's achievement, and the scores derived from these multiple observations have been demonstrated to measure certain achievements satisfactorily.

In other phases of educational research a similar state of affairs is found; the single observation has only the most limited value. If it is wished to obtain a socioeconomic index for comparing the background of children, it is unlikely that reliance could be placed on a single observation such as whether the child's home was or was not equipped with a telephone. While the presence or absence of a telephone is undoubtedly related to general socioeconomic conditions,

it is only one of many possible criteria. What would probably be done would be to prepare a check list of a number of items, each one of which is hypothesized to be related to socioeconomic conditions. The list might include items such as the presence or absence in the home of a telephone, a bathroom, an encyclopedia, a refrigerator, a separate bedroom for the child, and so forth. What would be done would be to add up the scores on each of these items by counting the checks according to a key. The total information provided by all of the items might be of use, while the individual items considered separately might be almost useless. The smallness of the amount of information encapsulated in each single item makes the use of many items necessary.

The reader should not conclude that there are no areas of science in which single observations are of great importance. Medical science is replete with examples of instances in which a single test provides almost certain knowledge about the presence or absence of a particular disease. Chemical analysis depends upon the making of a sequence of observations, each one of which is quite unambiguous in its interpretation. In physics too a single observation, such as is made in the determination of the density of a body, may provide highly valuable information that can be used to predict the behavior of the body in a multitude of situations. The contrast with educational research is marked.

The Combination of Observations

Once it has been recognized that observations must be combined in order to provide information that has any utility, two problems immediately arise. First, there is the problem of *what* observations are to be combined. Second, there is the matter of *how* they are to be combined. Most of the knowledge available about these two problems has been derived from studies of verbal responses to verbal problems such as appear in tests. Such knowledge, however, does have applicability to the wider range of materials used in educational research.

Consider the problem of estimating the size of a child's vocabulary or the relative size of the vocabularies of different children. One way of doing this might be to ask the child to define a series of words selected at random from a standard dictionary. If in this way we were to ask the child the meaning of ten words, and the child were to

define five accurately, we might infer that the child could define 50 per cent of the words in the dictionary. However, a random sample of ten words would provide only the most limited amount of information about the child's total vocabulary. Chance may have resulted in the selection of common words, or perhaps very rare words, or words that the pupil happened to know. If any of these things were to happen, our estimate of the pupil's vocabulary would be far out. In order to avoid this eventuality, several courses of action might be taken. One of these would be to use a large sample of words, the purpose of which would be to increase the precision with which it was possible to estimate the total vocabulary. The problem is the familiar statistical one of estimating the characteristics of a universe from the characteristics of a sample.

There is also a second way in which multiple observations may be used, which may be described by returning to the problem of measuring vocabulary. If we were to draw a sample of one hundred words from the dictionary by taking the last word on each right-hand page, or every twentieth right-hand page, it is probable that our list would contain many common words with which almost everybody was familiar. These words would waste time in administering the list and provide little information. Another procedure might be to select ten words that 90 per cent of the children to be tested would know, ten words that only 80 per cent of the children would know, ten words that 70 per cent would know, and so forth. This would form a rudimentary type of scale, which would be used to determine how difficult a level of vocabulary the individual can define. When such a scale is used, the purpose of measurement is no longer that of estimating the responses of the individual to the universe of possible items, although this can be estimated indirectly. It may be possible to infer that a person who scores seventy right on the scale can define in a similar way 90 per cent of the words in a given dictionary. However, the main purposes of such a scale are to compare one child with another so that deficiencies in vocabulary may be remedied, and to predict expected achievement in related fields.

Dimensionality and the Clustering of Observations

There are unsatisfactory features in the procedures discussed for the measurement of vocabulary which stem from the fact that knowledge of vocabulary cannot be considered to be a single unitary trait.

Consider, for example, the case of the student confronted with a vocabulary test consisting of equal numbers of scientific and nonscientific words. Such a test might contain fifty test items in each one of these two areas. Now it has been fairly clearly established that that knowledge of scientific words is not very closely related to knowledge of general vocabulary, and thus the test measures two rather distinct abilities. If the student taking the test obtained a score of sixty items right, it would be impossible to determine from the score alone whether the student had answered most of the scientific items right, or most of the nonscientific items right, or a considerable number of both types of items correct. Two persons might obtain equally high scores, one by obtaining a high score on the scientific items, and one by knowing the nonscientific items. The score alone would not indicate whether the person was strong in the one area or in the other. This is because the test does not consist of a homogeneous group of items all of which are measuring the same ability. If a measuring instrument is to have maximum utility, it should be designed so that it measures only one variable.

If a collection of items has such a mixed nature, then the score derived from it has only limited meaning. What is needed for meaningful measurement is a group of observations or items that belong together in some significant way and that can be used collectively as a measuring instrument. Such a group of items is sometimes referred to as a *homogeneous* scale, but one has to be careful about the use of the word "homogeneous." Some writers who talk about a homogeneous group of items mean only that the items all *appear* to be measuring the same kind of variable, as would be the case with a test of scientific vocabulary or a check list for measuring the socioeconomic status of the home. Others use the word to refer to a group of items that all belong together because it can be shown statistically that they all measure the same variable. What one really needs for meaningful measurement is a group of items or observations that not only belong together in some meaningful way but that also can be demonstrated by statistical means to measure a common property.

The problems of using multiple observations that have been considered up to this point involve the case in which it is possible to build up a measure to a point of usefulness by including additional observations. It was also pointed out that the observations added

must be homogeneous with those already available if the variable is to be truly meaningful. More commonly the research worker is faced with the problem of having at his disposal numerous observations that must be grouped together into separate and distinct scales in order to provide meaningful measures. It is this problem of grouping that must now be considered.

Combining Observations in Meaningful Ways

The scientist who approaches educational problems is commonly faced with an abundance of observations and must find some way of grouping them so that they provide useful information. In the conduct of many types of school surveys, such as are undertaken by school and college accrediting associations, the accrediting agency may accumulate large numbers of items of information about educational institutions. Such information is not easily handled as a mass, and in some way the observations must be combined into groups if they are to be easily interpreted. If the home backgrounds of children are being studied, the social worker or research worker may collect hundreds of items of information about each child. Such data may include items indicative of socioeconomic level such as home ownership, car ownership, size of home, value of home, and the number and type of appliances in the home. In addition, data might be obtained on a range of phenomena such as the number of brothers and sisters, the health of the parents, the sibling rivalries manifested by the children, the education of the parents, the number and type of books in the home, the preferences of the parents for the children, the father's occupation, the age of the parents, the number of neighborhood friends of the child, the religious affiliation, etc. If there are three hundred items of information collected about each of two hundred children, the resulting collection of sixty thousand items is not usable for most scientific purposes until it is organized in some way. Some of the ways of doing this must now be given brief consideration.

The a priori method. It is probable that the research worker who planned to obtain a large quantity of information about the background of the pupils had some theory about the characteristics of the background that were relevant for his purpose, which might have been that of predicting the level of academic achievement of the

pupils. He might have started out by postulating a number of different kinds of conditions in the background that might be related to achievement. One of these conditions might be the economic status of the home, another might be the degree to which tensions and frictions were absent from the home, another the cultural status of the home as indicated by the number of books or the presence or absence of a piano, and so forth. The list of items of information to be collected might well have been drawn up after a set of broad categories had first been established. After the data had been collected, the items would be grouped into these broad categories and a score derived for each. Thus, one score would indicate the relative socioeconomic status of the home, another the degree of psychological tension in the home, and another its cultural status. Thus, the three hundred items might be made to yield a dozen or fewer scores and the data would be reduced to manageable proportions. Advantages of the method are that it produces a set of measurements closely related to the theory on which the study was originally based and that the measures are quite likely to have considerable reliability.

While this method is attractive, it has its limitations, particularly in studies in which not too much is known about how the items of information should be grouped together. The latter is true of such an area as school characteristics, in which a check list might be used to describe a school by summarizing what are believed to be certain important facts about it. On the other hand, there are areas in which much is known about how items of information should be grouped. If the items referred to the biographical history of adults, it might be useful to group them in terms of the extent to which they reflected mechanical interests, scientific interests, clerical interests, and so forth.

The a priori method of grouping items of information rarely produces measuring devices that have particularly desirable properties as measuring devices. Too often the scales thus produced are too highly correlated one with another, which means that they measure characteristics that overlap. Considerable further work often has to be performed with these scales to refine them to the point where they are actually useful. A discussion of these additional steps is beyond the scope of this book.

Methods of grouping that depend upon the interrelationships of the items. The student is familiar with the concept that two tests are said to measure the same variable when they are highly correlated, and that they measure different variables when they are uncorrelated. The same concept may be applied to items of information of all kinds. On the basis of this concept, it is possible to sort a pool of items into groups on the basis of whether the items are or are not correlated. Each group of items would then include all those items that were highly interrelated from a statistical point of view. Also, the items in one group would tend to have little correlation with those in other groups—at least this would be so under ideal conditions. The actual procedure for doing this and the technical concepts involved are considerably more complex than is indicated here. Those who wish to explore the matter further are referred to an article on a special technique for doing this, called "the homogeneous keying techniques," which was developed by Loevinger, Gleser, and DuBois (1953).

Procedures for grouping items of information according to their interrelationships can be applied quite mechanically. Indeed, some of these can be undertaken almost completely by machine in this age of electronic computers. This is an advantage in terms of the speed with which the work can be performed, but the results are often difficult to interpret. Items that do not appear to belong together in any way are often grouped together. It is not that there are no good reasons why the items should be grouped in this way, but the reasons are not at all apparent. The reasons for the grouping often lie in the accidental way in which events happen together in our culture and in perhaps remote historical causes. For example, in one study of biographical material, it was found that three items grouped themselves together. These were (1) a knowledge of Latin, (2) a conservative political attitude, and (3) the ownership of a small business by the parent of the person being studied. It happened that in the city in which the study was undertaken, adherents to the Catholic faith were likely to share these three characteristics. This resulted in these three items being grouped together in a large number of biographical items tabulated by one of the statistical methods mentioned here; but without a knowledge of the community from which the data were derived, the grouping would be meaningless.

A second approach to the sorting out of items of information in terms of their relationships depends upon a whole series of techniques that have collectively become known as *factor analysis*. These techniques have evolved mainly for the purpose of attempting to identify the abilities that can be considered to underlie most of the aptitude tests that have been developed. Thus it can be shown that most of the numerous aptitude tests at present available can be considered to measure relatively few variables. One can classify tests in terms of the extent to which they measure each of a small number of reference variables, such as verbal ability, numerical ability, arithmetic reasoning, etc. By this means tests may be grouped together in terms of the *factor* that they measure. When large batteries of tests are used, say twenty or more, the procedure for conducting the factor analysis becomes arithmetically very laborious if it is to be conducted by hand, though modern electronic computers have enormously facilitated the procedures involved. There are also difficulties introduced by the fact that the procedures usually permit of more than one interpretation concerning the variables that are to be considered to underlie the battery of tests.

These same types of procedure can be applied to the grouping of test items or other items of information. One cannot point to examples where this procedure has been used with such items of information with striking success, but it is a procedure that is widely suggested in the literature. A major difficulty in its application stems from the fact that the arithmetic becomes extremely elaborate if the technique is applied to any large pool of items. Other difficulties arise in choosing, from the many possible sets of underlying variables that might be considered, the one that is most useful for the purpose at hand. At the present time the methodology of factor analysis has not demonstrated itself to be well suited to the sorting of items of information of the type considered, though it may have excellent uses in dealing with test variables.

The grouping of items of information in terms of the variables it is desired to predict. In a later chapter on problems of prediction, it will be pointed out that items of information are often assembled in order to make predictions. There are, for example, many studies in which information about the home and cultural background of children has been collected in order to predict and anticipate diffi-

culties in school. Let us consider such an example in order to illustrate the method of grouping items of information.

Let us suppose that items of information about the child's home and cultural background were collected for the purpose of predicting (1) reading disability, (2) social difficulties in school, (3) absenteeism, and (4) degree of success in an academic curriculum. The research worker collected 150 items of information about the background of each child entering a junior high school in a large city. The school served a residential area of varied economic circumstances. The data collection was continued until a record had been obtained of the background of each of 400 pupils. These pupils were followed through the school. A reading disability group was slowly and carefully identified. Reports that enabled the research worker to identify a socially maladjusted group were obtained from teachers and added to by counselors. Records of absenteeism were available, and so too were records of grades. Once the research worker had obtained all this information, he selected from the pool of background information items a group that predicted reading disability. When a total score was derived from this group of information items, the score was found to predict reading disability with considerable success in subsequent samples. A similar procedure was adopted for predicting each one of the remaining three variables that the study had been designed to predict.

While the procedure that has been just described does have a certain logic in it, there are procedures considerably more complicated that could be used to provide more accurate and efficient predictions. These other methods take into account not only the relationship of the items to the variable to be predicted, but also the extent to which the items provide overlapping information. This involves rather complicated statistical procedures, which really require the use of electronic computers if they are to be efficiently undertaken.

This method of using multiple items of information has been used successfully for combining items for all kinds of purposes. It has been used for predicting performance in various types of training programs, for predicting success in certain occupations such as that of salesman, for predicting delinquency, and so forth. In most studies the writer has seen that used items of background information, the resulting predictions have been extremely limited in accuracy. Of

course, this does not mean that the method might not provide highly accurate predictions when used to combine other types of information.

The disadvantage of the method lies in the possibility that the measures it produces, as in the example discussed, often are closely related to one another, and hence may all measure almost the same variable. One may suspect that three of the four scales produced may have all been measures of the extent to which the atmosphere of the home was favorable to intellectual development. The fourth scale, related to social adjustment, may have measured a different characteristic.

Some Cautions Regarding the Fractionating of Pools of Items

The procedures that have been discussed in this chapter are analogous to mechanical procedures for fractionating the various components of crude oil as it is brought up out of the ground. The crude oil is fed in at one end of the fractionating plant, and a whole range of petroleum products, some useful and some not, comes out of the other. Procedures for grouping and selecting test items are of this character. They can be applied and used without much knowledge of the why and the wherefore, and their application can be considered in most cases as a step in the direction of developing tools for research, but it can hardly be considered to be research.

This point is made because the writer has frequently been faced by the graduate student who has come to him with the proposal that his doctoral dissertation consist of the application of a certain technique to a pool of items that the student proposes to build. This is not an activity that should be encouraged on the part of the graduate student. It represents a laborious and time-consuming routine. It does not encourage the type of activity often considered to lie at the very core of a program of doctoral studies—namely, the thinking through of an important problem to the point where ways are found of arriving at a solution. Research of any consequence requires more than the mere application of a mechanical routine. If it could be undertaken by the latter means alone, it could be produced in a factory by a relatively uneducated labor force.

Much of what is undertaken in the name of research, and which is the wholesale application of a technique to some universe of items,

fails to note that the mere identification of a variable is no guarantee that it is going to be of any use. As a matter of fact, new measuring instruments in the behavioral sciences can be developed very easily. The great difficulty is to discover and develop variables that have predictive value. This usually requires prolonged endeavor and what has been termed scientific insight, and does not result from any mechanical procedure, but rather does it come from the development of a sound theory and the testing of deductions from this theory.

It thus behooves the scientist to begin all work on the development of measuring instruments with a theory concerning how the variables he is attempting to measure relate to the specific aspects of behavior that he is studying. If this is done, it will be necessary to test the instruments in order to determine whether the measures they provide permit the making of the predictions that were anticipated. If they do not, one should discard not only the instruments but also the theory on which they were based. Measures that have merit in terms of their internal properties do not help in building a sound theory and do not contribute to knowledge if they show no signs of operating in the way expected.

Finally, it must be reiterated that the information provided by a single observation related to education is extremely limited, and even a group of relatively homogeneous items cannot be expected to provide more than a hint about other types of events. This may be looked upon in another way in the case where the observations refer to behavior. If a group of test items occupies five minutes of time, this must be recognized as an extraordinarily limited sample of a person's behavior. From what a person does in a five-minute period, there is only the most limited ground for generalizing about what the same person will do during other five-minute periods. At the same time it should be recognized, of course, that a test situation is not *any* sample of behavior, but should be a sample of behavior that has been demonstrated to have particular significance for predicting how the person will behave in certain other situations. Nevertheless, there is a definite limit to the amount of information that can be obtained about behavior in a given period of time.

Finally, a word of caution must be said concerning the development of tests that cover a very narrow range of phenomena. Such

tests, because they cover a very limited range of phenomena, can be used to predict only a very limited range of phenomena.

There is something of a paradox here. The more specific the variable that is measured, the greater are its chances of being independent of other variables, but the less are its chances of being a measure that has wide utility as a predictive device. It may be remembered at this time that the measures that have had the most widespread use in predicting behavior are those that tend to be rather generally correlated with a wide range of tests. For example, verbal-factor tests have the most general utility for predicting trainability, and yet these tests show considerable correlations with other tests. It is likely that the fact that they correlate with other tests is, in itself, symptomatic of the capacity to predict.

The items considered must cover a range of activity in order that the resulting measuring instrument may apply to a range of activities. If one postulates that liking for mechanical activities and objects clusters, it would be undesirable to limit the range of mechanical objects and activities included in a test to those involved in the repair and maintenance of automobile engines. Such narrowness of range is unlikely to be of use for most purposes for which the instrument is to be used. The most striking failure of this kind in the development of instruments is where the same question is repeated several times within the same test. One test that attempts to measure needs is of this character. What such a procedure does in this case is to build up reliability for each measure of need, since when approximately the same question is repeated, it is going to be answered in approximately the same way. However, what this does is to measure a highly specific characteristic, which may have little generality in other situations in which it may be expected to appear. On the other hand, it is this specificity that will ensure that the various measures of need are independent of one another.

The result is that the instrument appears to provide a series of measures of need that have both high reliability and high independence of one another. This is generally considered to be a most desirable state of affairs, but it is not so when it is achieved by measuring responses to extremely restricted situations. Exactly the same would be true in dealing with other kinds of observations. If a check list were made to be used for obtaining a measure of the adequacy

of a school plant, it would be undesirable for most of the items to refer to the size and adequacy of the library. There would be agreement that a score to be derived from the check list would be useful only if it provided a comprehensive coverage of the school facility.

SOME SPECIAL PROBLEMS OF UTILIZING MEASUREMENTS

Measures in Which the Responses Are a Function of Time

The measures that have been discussed to this point are those customarily administered in such a way that increases in the time available for responding would not appreciably change the score. This is inevitably the case in most tests of information, knowledge, thinking skill, attitude, interest.

Test instruments that involve speed of response do not involve the problems of item selection that have been discussed in the previous sections of this chapter. The reason is that speeded instruments are almost invariably such that the score is the number of units completed in a given time, and it is important that each unit be equivalent to every other unit. A common type of speeded test is the clerical aptitude type, which requires the individual to compare two lists of closely similar names and to mark those where corresponding pairs of names do not match. Each pair of names corresponds to a unit of work, and the test should be so designed that all pairs are equivalent as units of work. In order to approximate equivalence, the person making the test will need to have some theory concerning the factors in these stimuli that are related to behavior. He may, for example, believe that long names are more difficult to compare than are short names, and hence, in order to make all units of work equivalent in terms of this factor he will make all names equal in length.

It may be pointed out here that when time is used to control a test score, the method commonly used is that of controlling the time on the total test. Customarily, all examinees start and finish at the same time. Under this procedure, a person who works at the maximum speed of which he is capable and completes ten items obtains the same score as the person with facility for the task who works along in a way that is leisurely for him and also completes ten items.

A superior way of controlling the speed factor in such tests would be to expose each item separately for a given interval of time. At the end of that time, the next item would be exposed, and so forth. In this way, the examinee would be paced all through the test. The time per item might stay constant under these conditions or might be reduced as the test proceeded.

Pattern Analysis as a Method of Combining Observations

The discussion of the problem of combining observations that has been presented up to this point has disregarded a problem that clinical psychologists have frequently stressed. It is that the pattern of responses to particular situations may provide more information than a simple summation of those responses. Problems of pattern analysis will be first discussed here with reference to the problem of combining and utilizing scores for different tests, since such problems are already familiar to the teacher.

Rorschach administrators have commonly stressed that it is the interrelationship among the various scores on the instrument that provides information of real significance, and that the absolute values of the scores mean little. As far as the writer is concerned, he knows of no cases where it has been clearly demonstrated that the patterns of scores on the Rorschach, rather than the scores themselves, are of real significance. There are, however, cases with other types of measures where substantial evidence has been collected to show that patterns of scores may be extremely important as predictors. In some of French's studies (1956), effectiveness in particular situations has been predicted well in terms of the relative strength of affiliation and achievement motives. In such studies, situations have been presented that produce a conflict between affiliation and achievement motives, and the action that results is a product of primarily only one of these motives. As one might expect, it is the stronger of the two motives that ultimately becomes the major determinant of behavior in these situations. In such a situation, three patterns of motivation are possible. Either achievement motivation is the stronger of the two, or affiliation motivation is the stronger, or they are both equal in strength. If a third motive were involved and each motive were considered to assume three levels (high, middle, and low), then twenty-seven possible patterns of motivation may occur.

It can be readily seen that if as many as 6 motives are involved, the number of possible patterns becomes large, and if data are to be collected with as many as 100 cases showing each pattern, then very large numbers of cases are to be collected, with a minimum of 62,900. In practice, far more than this minimum would have to be collected, since there would certainly be far more cases in some categories than in others. Perhaps 1,000,000 or more cases might have to be collected before there were at least 100 in each category. It is the cumbersomeness of the data to be collected that makes profile analysis unsatisfactory where many possible patterns exist. It is for this reason that some restrictions must be placed on the process.

One method of reducing the number of patterns is to use only a high-low dichotomy instead of a high-middle-low division. Another is to group together patterns that can be considered similar on some rational basis. While this latter approach is attractive, there are difficulties involved in developing a rationale that can be used for the grouping of patterns. The writer is inclined to believe that comprehensive studies of patterns in which every pattern is studied in relation to a criterion are unlikely to provide useful results. It is too much of a hit-or-miss procedure. However, studies of the patterning of a few variables for which there is strong reason for believing that differences in pattern are associated with differences in performance may well prove to have value. Here again the writer is attempting to stress the need for well-defined hypotheses or problems before data are collected or analyzed. Massive pattern analyses conducted in the hope of finding a vague "something" run counter to this concept of research.

The problem of pattern analysis becomes important in the scoring of specific tests when it is desired to produce a score that will predict a particular criterion with maximum accuracy. Meehl (1950) was the first to see the real significance of this problem when he pointed out that two dichotomously scored items could each correlate zero with a dichotomous criterion, and yet from the pattern of responses to the two items it might be possible to predict the criterion with perfect accuracy. This phenomenon is illustrated in Table 2. It is clear in this table that two patterns Yes-No and No-Yes predict the criterion, and two patterns Yes-Yes and No-No predict failure. It is also shown that each one of the items considered separately has no value

at all in predicting the criterion. It can be shown that what is happening under such circumstances is simply that a curvilinear function is being used for making the prediction, instead of the linear function that fails to predict.

The Meehl phenomenon is a special case of the proposition that scoring a test in terms of patterns provides the best possible prediction of a particular criterion. Scoring in terms of number right may provide as good a prediction in some cases, but it cannot provide a better prediction than when all possible patterns are used in making the prediction.

TABLE 2. Patterns of Responses to Two Items in Relation to a Criterion

	Response on Item I	*Response on Item II*	*Number in Each Pattern*	*Pass or Fail*
Pattern I	Yes	No	25	Pass
II	No	Yes	25	Pass
III	No	No	25	Fail
IV	Yes	Yes	25	Fail

Relation of Each Item to the Criterion

Item I	*Pass*	*Fail*
Yes	25	25
No	25	25

Item II	*Pass*	*Fail*
Yes	25	25
No	25	25

Ordinary scoring methods in which all persons obtaining the same score (say the number right) are placed in the same category—even though they answer different combinations of items—present a case in which many patterns are classified and grouped together. The conventional method of scoring is likely to be less efficient than would be a scoring system which kept all patterns separate.

In the case of certain types of scales, known as Guttman scales, only a limited number of patterns are possible. Consider, for example the following five-item scale:

1. $4 + 6 =$
2. $3\frac{1}{4} \div 2\frac{1}{2} =$
3. If $2x + 4 = 0$, then $x =$
4. If $x^2 - x - 12 = 0$, then $x =$
5. If $y = 3x^4 + 2x^2 + 4$, then $dy/dx =$

In the case of this five-item scale, the person who was able to answer the last item would, almost certainly, have been able to answer the previous four items. The person who answered the first two items, but who was unable to answer the third, would almost certainly fail on the last two. The scale is so graded in difficulty that a person is able to answer the items up to a certain point and then fail all items beyond that point. In the case of such a five-item scale, only six possible patterns of response are possible, and these would be as follows:

	Item I	*Item II*	*Item III*	*Item IV*	*Item V*
Pattern I	Wrong	Wrong	Wrong	Wrong	Wrong
II	Right	Wrong	Wrong	Wrong	Wrong
III	Right	Right	Wrong	Wrong	Wrong
IV	Right	Right	Right	Wrong	Wrong
V	Right	Right	Right	Right	Wrong
VI	Right	Right	Right	Right	Right

In such a case there are only six possible patterns. The score can vary from 0 to 5, and from the score alone it will be possible to determine the pattern of responses that produced that score. If the test did not form such an ordered scale, other patterns of responses would be possible, and in the extreme case where no semblance of an ordered scale were present, then there would be thirty-two possible patterns of scoring categories. Only a very few achievement tests and attitude scales form ordered scales of this type. Most types of measuring instruments are so remote from being ordered scales that numerous patterns of responses are possible. A method of considerable interest for scoring these various patterns has been suggested by Lubin (1957), who has done much to clarify concepts in this field.

The way in which Lubin proposes to score patterns can be shown to be the most efficient of all possible methods of using the information that the test or other device can make available. In order to explain how this is done, let us consider a relatively simple example. Suppose that a measure were available of the satisfaction that teachers derived from their work, and it was desired to relate this measure to a five-item personality test that had been administered at an earlier date. In actual practice, one would not think of utilizing such a short

test, because so few items would contain relatively little information, but a five-item test is convenient for the present explanation of Lubin's method of scoring, which is named *configural scoring*. The average score on the job-satisfaction scale for all those who have the same pattern of responses on the personality scale is the score for that pattern. Thus for each one of the thirty-two possible patterns a score on the job-satisfaction scale will be assigned and this will be the configural score. It can be shown that the configural score is the score that will best predict the job-satisfaction scale from the personality scale. No other method of scoring can provide a better prediction, though, of course, it is quite possible that the personality scale may be a poor predictor of job satisfaction in the teaching profession.

In the case of a 5–item scale, there are 2^5 possible patterns of response if there are only two ways of answering each item. In the case of a 10–item device there would be 2^{10} possible patterns, which is 1,024. In the case of a 15–item device there are 32,768 possible patterns. This points up the real difficulty in using this type of pattern analysis. There are likely to be just too many patterns to be manageable if there are more than a few items to be considered.

Just how much is to be gained by pattern analysis in the use of multiple observations? No answer can be given to this question at the present time, since experience with such methods is still limited. It is quite possible that if studies are conducted where there is real reason to believe that patterning has important effects, real gains may be found in the use of this type of method.

Reliability of Measurement

Many of the operations discussed in this chapter that involve the combining of observations are undertaken for building up those characteristics of instruments that in the history of measurement have been referred to collectively as *reliability*. Problems of reliability refer mainly to a special class of inferences from scores. In the history of psychological measurement, a reliable measure has been considered one that would remain stable if the measure were again applied under similar conditions. This statement implies that differences in scores on the measure from one person to another are not merely the product of a great number of uncontrolled events, but that they represent some relatively stable and continuing condition that differentiates

persons one from another. For example, absenteeism in any school on February 8 of any year may show great differences from school to school, but these differences are a product of a multiplicity of causes. In one community a high absentee rate is due to a local epidemic of measles, in another it is a result of a blizzard, and so forth. We may expect little relationship between absenteeism on February 8 and absenteeism on March 8 of the same year. Absenteeism on a particular day may be said to lack sufficient reliability to make it a useful measure for any conceivable purpose. A measure that is to have value must be determined by conditions that have some permanence and continuity, and it is this that in turn gives the measure stability.

A central weakness in the whole concept of reliability stems from the difficulty of defining what is meant by *similarity of conditions.* It is quite obvious that if measurements were repeated under truly *identical* conditions, the results would inevitably be identical. Measurements vary because conditions vary. The expression *measurement under similar conditions* cannot refer to *measurement under identical conditions,* but just how far conditions can depart from identity and still be considered similar is entirely a matter of personal judgment.

Reliability is thus a somewhat fuzzy concept. At least a part of the fuzziness is a result of the fact that the term refers to a series of concepts that are confused with one another. Thus the American Psychological Association (1954) has wisely suggested that any manual that accompanies a test and that provides estimates of reliability should indicate the method by which it was computed. Different estimates of reliability pertain to different inferences.

In a split-half type of reliability, scores derived from half of the items are correlated with scores derived from the remaining half. If the items are considered to be random samples of a universe of items, and if they are divided into two sections at random, then the reliability coefficient is an attempt to answer the question, "To what extent is it possible to make inferences from a score on one random sample of these items concerning scores on another random sample?" However, if the two groups of items are so matched that for each type of item in the one group there is a corresponding item in the other group, then the inference pertains to the extent to which a score

from one sample of items covering certain specific areas can be used to infer scores on other similarly structured samples.

In the case of the coefficient of reliability based upon two successive administrations of the same test, the purpose is to estimate the extent to which it is possible to infer scores at other points in time from a test score obtained at a particular time. It is a rather different inference from that made from a split-half or a parallel-form procedure for estimating reliability.

The estimation of reliability by means of the Kuder-Richardson type of formula, which has already been mentioned, refers again to a different type of phenomenon. Cronbach (1951), who has made a careful study of this approach, refers to the coefficient derived from this procedure as *alpha* rather than by the more lengthy name that it has acquired from its originators. Cronbach has also shown that it refers to an internal property of a test, which is a product of the statistical relationship among the items. This property is known as *homogeneity*, and refers to the extent to which all the items on a test can be considered to contribute to the measurement of a single common variable. This is stating the matter in the simplest possible terms. A precise definition of homogeneity requires the extensive use of mathematical terms. The coefficient *alpha* is for this reason now most commonly referred to as a measure of homogeneity rather than as a measure of reliability.

Measuring instruments of the type used in the social sciences can provide only measures that approximate homogeneity. Much of the variance of each item can be attributed to sources other than that which it is desired to measure. However, this may not be as harmful to the total score as one might at first assume. The reason for this is that the unwanted aspects of the variance are derived from a large number of unrelated sources and therefore, so to speak, tend to cancel out one another in a total score.

Summary

1. In educational research it is usually true that a single observation provides only a very limited amount of information. In order to overcome this limitation, it is usually necessary to combine together several observations.

2. Items of information need to be grouped together in some way so that they form a meaningful measuring instrument. Under ideal condi-

tions the items should all belong together in terms of some theory, but they should also all belong together in a statistical sense, and in this sense should all measure a common variable.

3. Items of information may be combined together on the basis of judgment either because they appear to belong together or because there is some theoretical basis for grouping them together. This method of grouping together items of information is referred to as the a priori method.

4. A series of methods of using multiple observations have been developed in which the grouping depends upon the statistical relationships among the items. Two general classes of techniques for this purpose have been developed:

> a. The homogeneous keying technique represents a method that has been evolved specifically for the purpose of sorting a large number of items of information into groups each of which forms a measuring instrument.
>
> b. A second approach is that of factor analysis, which evolved more as a procedure for the grouping of tests than a procedure for the grouping of items.

5. Observations may be grouped also in terms of the extent to which they predict some other variable.

6. Caution should be exercised in the combining together of items of information. The procedure should be so planned that the resulting variables are meaningful in terms of current educational theory. In any case, the amount of information that one may expect to obtain from even a group of items that form a measuring scale is limited. The more specific the variable that is being measured, the more it is likely to be independent of other variables and the more it is likely to predict only a narrow range of behavior.

7. Pattern analysis represents a special group of techniques for using multiple observations. While clinical psychologists have long believed that the pattern of scores on a battery of tests might be of greater significance than the actual values of the scores themselves, it is only recently that a theory of pattern analysis has been developed. Pattern analysis can also be applied to the scoring of items, and it can be shown that configurational scoring is the most efficient method of utilizing all of the information provided by a set of observations.

8. Pattern analysis techniques are still being explored. Those that are available tend to be extremely cumbersome to use. As yet, it cannot be stated how much is to be gained by pattern analysis techniques as contrasted with simpler and more traditional techniques.

9. The development of methods for using multiple observations rather

than single observations has been intimately related historically to the problem of improving the reliability of measurement. As understanding in this area has developed, the concept of reliability has been found to be more and more unsatisfactory, and those who have been engaged in the development of measuring devices have been urged to specify just what technique they have utilized in the estimation of reliability. The concept of homogeneity seems to be a more satisfactory one than the concept of reliability.

Validity of Measurement | 7

A central focus of discussion, whenever problems of measurement are considered in education, is the concept of *validity*. Unfortunately, this concept has drifted into the behavioral sciences by an adaptation of the word "validity" as it is used in common speech. Had it been introduced by some careful thinker at an appropriate point in the history of the development of the behavioral sciences to denote some well-defined concept, the difficulties that have occurred over the past forty years in clarifying its meaning might not have arisen, but such was not the case.

The present state of thought concerning the problem of validity is probably best understood by reviewing the history of the concept. The struggle for clarification that psychologists have lived through during the past forty years has resulted in much insight and understanding, and the historical review is presented here in order to bring some of this to the student.

Early Attempts to Standardize Measurement of Behavior

The American Psychological Association has had a long-standing interest in the standardization and use of psychological tests. As early

as 1895, the Association appointed a committee on mental and physical tests and reported its recommendations at the annual meeting held at Boston in 1896. A more ambitious venture was started a few years later, when in 1906 the Association established a committee on the subject of measurements. The main purpose of this committee seems to have been the collection of descriptions of tests in use, in order to make them available to other investigators. It was felt that if the results of different experiments were to be comparable, it was necessary for investigators to use at least similar instruments of measurement. The object of this entire procedure was to facilitate the development of generalizations concerning the relationship of test variables to other variables. The implication in much of the report of this committee was that the tests defined the variables that were measured, but the term "validity" was not used, and nothing seems to have been lost by its absence, for the tests discussed in the report were discussed within the framework of a surprisingly well-developed rationale derived largely from the associationistic and Wundtian tradition. Not all modern test batteries have such a rich background of theory as was the case with this battery proposed for general use. However, the excellent work of this committee was ahead of its time, for it is hard to find a single reference to its work in the technical literature of the period.

Psychologists contemporary with this committee continued to develop large numbers of measuring instruments. One of the most prolific of these was Edward L. Thorndike, who in his *Theory of Mental and Social Measurement* (1904) put forward the view that psychological measures represent facts about an individual, and that the problem of the psychologist is to arrange conditions of measurement so that the measures are accurate representations of the facts. In this connection Thorndike uses the example of a person reacting to a stimulus. From this behavior, the psychologist may abstract the quality of speed of reaction, and the problem of measurement is to devise an instrument through which a number that will represent the reaction time can be assigned the individual's reaction. Refinements of the measuring procedure merely reduce various errors that contaminate it. It should be noted that there is a real difference between the argument that measures are an attempt to describe accurately certain physical events such as reaction time, and that measures are an

attempt to represent some underlying and forever unobservable psychological reality. Arguments about "what tests really measure," which for twenty years were considered to be the central problem in determining whether tests were or were not valid, arose at a later date. In Thorndike's subsequent work this problem is raised, and his *Measurement of Intelligence* (1927) opens with the statement that psychological measuring instruments still have three fundamental defects, which are: (1) "Just what they measure is not known"; (2) "How far it is proper to add, subtract, multiply and divide, and compute ratios with the measures obtained is not known"; and (3) "What the measures signify concerning intellect is not known." The first of these three implies that there is an underlying reality, measured by tests, that can be known, and so too does the third. However, in the works of Thorndike up to and including his 1927 volume *Measurement of Intelligence*, the term "validity" is used only with respect to the validity of judgments of the difficulty of tasks. This is true also of his *Educational Psychology* (1913).

The first use of the word "validity" in a technical article, as far as the present writer can determine, is in an article by Freeman in 1914. In this article, Freeman states that "this report deals only with questions regarding the technique and validity of test methods." This statement implies that the term "validity" was then used in discussions of testing problems, even though it was not commonly used in the literature of the day. However, Freeman does not use the term again in the remainder of his article. Just a few years later, the discussion of the concept of validity became involved in discussions of what tests really measure, and a valid measure was defined as one that measured the variable it was supposed to measure.

Discussions of *what tests really measure* and whether they do or do not measure what they are supposed to measure waxed in the early 1920's, largely as a result of the widespread application of intelligence tests. Discussions of the subject appeared not only in technical journals but also in popular magazines. The disillusionment concerning the value of these tests, which inevitably followed an era of excessive and premature enthusiasm, is well represented by a series of articles on the subject by Walter Lippmann, which appeared in the *New Republic* in the early part of 1923. Most of what he had to say would conform to modern thinking on the subject. The articles prob-

ably had some desirable effect of inhibiting those who made immoderate claims concerning test usage or who claimed that tests measured innate faculties. A more constructive article followed in the June 6, 1923, issue. It was by Edwin G. Boring, who presented the view that today would be identified with the operational point of view, namely that "intelligence as a measurable capacity must at the start be defined as the capacity to do well on an intelligence test." The implication is that it is not reasonable to ask the question, "What do tests really measure?" Measures derived from tests represent certain characteristics of behavior that have been selected because according to some theory, they have a special value for prediction. Tests measure whatever they measure. They are valuable if they are capable of making the predictions they are expected to make according to the theory on the basis of which they have been developed. Only empirical verification can show whether the predictions hypothesized can be made. The discussion of what tests really measure should have been settled then and there, but it has continued intermittently over the years.

In the late 1920's and in the two decades that followed, the profitless discussion of what tests really measured became displaced by a concern for the problem of what inferences could be made from test scores. But the psychologists both in education and in industry did not state the problem in this way, for their interests were strictly practical.

This new trend was a result of the growth of applied psychology, and the interest of applied psychologists was largely in the matter of what tests predicted. In this context, psychologists began to refer to correlations between test scores and variables that it was desired to predict as *validity coefficients*. The acceptance of this concept of validity was more a matter of convenience than it was a product of profound reflection. While nobody cared too much any more what a test measured, it became a matter of paramount importance to determine whether a test was or was not capable of being used for the making of predictions. This aspect of the matter has been the central focus of attention when the meaning of the term "validity" has been discussed during the last twenty years.

Since evidence of validity for a test became more and more a guaranty of a good market, test manuals in the 1930's and 1940's showed

an increasing liberty with the use of the term, and *evidence for validity* became progressively more and more remote from evidence that useful predictions could be made with the device. A situation rapidly developed where the many new meanings of "validity" had to be defined.

In the late 1940's, voices began to question the clarity and precision of the then current concept of validity. Mosier (1947) wrote a penetrating article pointing out that the term "validity" was used with reference to four somewhat distinct concepts: (1) validity by assumption, where the items of a test appear to bear a logical relationship to the phenomena to be predicted; (2) validity by definition, where a test is used to define a particular variable, as would occur if knowledge of mathematics were defined in terms of a score on a particular mathematics test; (3) face validity, which occurs when a test, in addition to having statistical validity, also appears to have relevance; and (4) validity by hypothesis, which occurs when the mass of previous evidence supports the contention that a test has relevance for predicting a particular criterion.

Mosier's article heralded a series of attempts to clarify concepts in this area. The most ambitious of these attempts must now be considered.

The American Psychological Association's Second Attempt to Order Concepts in the Measurement Domain

As a consequence of the failure to reach any kind of agreement among professional persons concerning the meaning of validity, each test publisher felt free to interpret it in his own way, and often in a way that was at variance with a large fraction of the profession. Other concepts in the measurement field also lacked clarity but perhaps were not obscure to the same degree.

The proliferation of instruments that has filled the markets during the last two decades, the varying standards that have been adopted by test publishers, and often the lack of standards, made it desirable for the American Psychological Association in 1949 to take some action in setting up standards that test publishers might follow. The committee, which became known as the Committee on Test Standards, spent the first few years of their efforts on the preparation of a report, which has particular relevance here because of its novel

contribution to the development of the concept of validity. It is understood that the original committee was in favor of discarding the word "validity" but decided to retain it because its usage had become so deeply engrained in the field of applied psychology. The thought seems to have been that the concept could be discarded and replaced at a later date. In the interim period, it was decided to define various aspects of validity and to name them separately. These various meanings need now to be considered.

Predictive validity. This is validity in the customary sense in which it has been used in applied psychology and in aptitude measurement in education. It is validity as represented by statements such as, "The Jones Reasoning Test administered in high school correlates 0.4 with average grades over the first two years in liberal arts colleges having enrollments greater than one thousand students." Such statements represent empirical relationships that supposedly can be used to evaluate the worth of a test for a particular purpose. In the statement that was just quoted, the data included—which, let us assume, are unimpeachable—permit us to make statements concerning the value of the instrument in the particular situation in which they were collected. They do not provide knowledge concerning how the test will work in other situations. If inferences are to be made concerning the predictive validity of the test in other situations, it is necessary to make assumptions concerning the relationship between the situations in which the data were collected and new situations in which it is desired to make a prediction. Usually very little knowledge is available concerning this relationship. Indeed, in spite of guesses to the contrary, an investigator might be quite surprised to find that the Jones Reasoning Test did not provide as satisfactory predictions within teachers' colleges as within liberal arts colleges. *Predictive validity really does not provide a basis for using an instrument except in the situation in which it was validated.*

Concurrent validity represents a concept very similar to that of predictive validity. The difference is a relatively minor one. In the case of predictive validity, the variable used in prediction is measured at some time previous to the measurement of the variables that are predicted, as would be the case in predicting college grades from aptitude tests administered in high school. On the other hand, in the case of concurrent validity, performance on a test is compared with

a measure derived from some concurrent performance, as when a test given in high school is used to predict high school performance. The fact that a test has concurrent validity does not necessarily mean that it has predictive validity. For example, it is quite conceivable that a test might discriminate in a certain plant between those who had and those who had not been promoted, but it is quite possible that the same test given when persons were first hired might have no success at all in predicting which would be promoted later. The test might involve only information acquired after the person was hired.

Real and important questions can be raised concerning the merit of separating concurrent and predictive validity. When predictions are made over any time interval different from that used in the original validation of the test, assumptions must be made concerning the justifiability of the predictions under the new conditions. Sometimes the assumptions are reasonable and sometimes they are not.

Content validity is the extent to which the situations included by the test are representative of the group of situations that the test is supposed to sample. It is common in the case of achievement tests to compare the content of the test with the content of the curriculum and to arrive at some judgment concerning the relationship of the two. The product of such an operation is a rough-and-ready judgment, and no satisfactory methods have been devised for quantifying this relationship. The central difficulty involved is that of measuring the characteristics of situations so that a sound basis exists for comparing the properties of the test situation with the properties of the situations that they supposedly represent.

The difficulty of measuring content validity reflects a current inadequacy in perhaps all theory in the behavior sciences. Until it is possible to measure the characteristics of situations to which persons respond, there is little hope of obtaining in psychology generalizations that have really broad significance. If the laws of behavior in situations having characteristics X, Y, and Z are known, one may be justified in making predictions about behavior in other situations having the same characteristics X, Y, and Z—but only if it is genuinely possible to identify these characteristics.

Construct validity is the final category proposed by the committee of the American Psychological Association. Construct validity is demonstrated by showing that measures derived from the instrument

can be used for making inferences consistent with the theory on which the test is based. Thus a projective test of achievement motivation is found to provide scores that are correlated with output of work on a task so simple that output must be considered to be mainly a function of motivation. If this were found to be the case, one could reasonably infer that evidence had been elicited to support the contention that the measure could be identified with achievement need. Usually it would require more than the single piece of evidence of the type studied to justify the inference about the characteristics of the variable measured. What is needed is evidence from remotely differing spheres, all of which substantiates the belief that the instrument predicts in those situations in which prediction is expected. The nature of science is such, of course, that one cannot expect all of the evidence to point in the same direction. One can expect some inconsistencies, and it is the functon of further work to discover reasons for these inconsistencies and to revise the theory under review so that it becomes consistent with all available evidence.

Sometimes the evidence that is dealt with in the determination of construct validity is derived by correlating scores from the instrument under consideration with scores derived from similar tests, or tests that are related according to some theory. It seems to the present writer that such evidences are necessarily weak and limited in value. Mere relationships among measuring devices are limited sources of information. What is needed is relationships between measures derived from devices and important variables that one wants to predict.

It should be noted that other writers had discriminated the concept of construct validity from the broad and hazy general concept of validity. Mosier's "validity by hypothesis" (1947) is essentially the same. Gulliksen's term "intrinsic validity" also carries with it much of the same meaning. As a matter of fact, as far back as 1936, Bowers (1936) clearly defined the concept of construct validity and pointed out that it was the only real basis for generalization.

An Attempt to Restate the Problem

The present attempt to restate the problem in terms of the general matter of scientific measurement stems from the observation that the behavioral sciences are the only sciences that seem to need the concept of validity. Furthermore, in the field of behavior, it is only

within the limited domain of testing that the word "validity" is used. Experimentalists, theory-builders, and clinicians do not seem to feel any need for the concept in their respective domains. Works are written on the basis of experiments, and theories of psychology are drawn up without reference to the term. Why do those in the testing field need this concept that researchers in other fields of psychology and in other fields of science find superfluous?

The problem of validity seems to deal with the problem that in other fields of science is known as that of generalization and inference. When the psychologist asks the question, "Is this test valid for this purpose?" he is asking the question, "From the response of the individuals to this test, what statement can be made concerning the response of these individuals to this other situation that by custom is called the criterion situation?" The question concerns what inferences can be made from test scores.

The concept of validity, as it is commonly discussed, refers to the problem of inference and generalization. Discussion usually involves a test score R_t from which it is believed that some response in a criterion situation R_c is to be predicted. Validity is the extent to which it is possible to base statements about R_c on R_t. Sometimes it refers to the extent to which it is possible to make generalizations from R_t to R_{c1}, R_{c2}, R_{c3}, or a whole class of criterion situations.

Since the usual evidence that justifies such generalization and inference is a correlation coefficient, the value of this evidence must be given some consideration at this point.

Correlation and Inference

The fact that the Jones Reasoning Test has been demonstrated to correlate with the grades of students in, say, an elementary electronics course is extremely limited information, which permits little or no generalization. The fact that the correlation between test scores and grades was found to be 0.4 ($N = 200$) on a single class cannot be taken as a sound basis for making the generalization that the relationship will be maintained in future classes. The correlation may have generated by some incidental condition, which may rarely or never recur. The author can recall one case in which a correlation between an aptitude test and a set of grades was generated by the fact that the instructors reviewed the aptitude scores of their students

just before they assigned grades. In other words, the mere fact that a correlation has been found between an aptitude test and a criterion variable is very inadequate evidence that any generalization can be made concerning the probability that the same relationship will be found on future occasions; or, in terms of the traditional language of psychology, a single correlation coefficient, even if it is of substantial magnitude, is poor evidence of validity. Predictive validity and concurrent validity are really quite trivial concepts, because the correlation coefficients on which they are based are insufficient for establishing a generalization that can be used.

If a correlation coefficient does not permit useful generalization concerning the value of a test, then what does constitute such evidence? *The primary evidence on which useful generalization can be based is the fact that the relationships among the variables are to be expected on the basis of a more general theory of prediction that has been shown to have value in the making of predictions.* For example, consider the problem of constructing a test for reducing the percentage of failures among those admitted to law school. Suppose that an experimental battery for solving this problem included a test consisting of the well-known cube-turning items * and a verbal reasoning test that involved a considerable knowledge of vocabulary. Suppose that it were found that the reliability of each one of these two tests was 0.95 for an entering class of 230 law students, and that the space test and verbal reasoning test correlated 0.5 and 0.3 respectively with the average grades of this group while in law school. For convenience, let us assume that there were no drop-outs. In addition, let us assume that the data show that the cube-turning test is likely to provide as satisfactory a prediction of average grades as an optimum combination of scores from the two tests.

In such a situation, the choice of the empiricist would be clear. He would use scores on the cube-turning test for the selection of the next entering class, for had not this test been established as the *more* valid of the two? In addition, he would reflect that a statistical test of significance had demonstrated to his satisfaction that the difference in the "validity" of the two tests should not be considered as a mere product of sampling.

* Cube-turning items usually illustrate a sample cube with a different design on each visible surface. A number of other cubes are illustrated, and the subject must select from among these the one that is the same as the sample cube but turned into a different position.

However, serious questions may be raised as to whether the choice of the empiricist is a sound one. In criticizing his choice, the scientist would point out that previous studies had shown with monotonous consistency that the type of verbal test used in this study was a more satisfactory predictor of grades in courses involving the extensive manipulation of verbal symbols, and that a space-manipulation test had been shown to be notoriously inadequate for this purpose. The fact is that the data collected in this situation are so inconsistent with data previously collected, and with generalizations based on those data, that no acceptable generalization is possible concerning the extent to which these tests are likely to be generally useful for selecting law students for future classes in the same law school or in other law schools.

Most persons familiar with the generalizations that can be made concerning the value of various types of tests for the selection of students would probably agree that the verbal test would be more likely to have selective value for law students than the space test. The verbal test can be considered to have "validity" for the selection of law school students, not just because of the single empirical finding of $r = 0.3$ for this particular case, but because this finding is consistent with an organized body of knowledge. No generalization seems reasonable from the so-called validity coefficient of 0.5 for the spatial test, simply because this coefficient is inconsistent with a substantial body of available knowledge and thus, in traditional terminology, cannot be considered valid as a selection instrument in this particular situation.

The moral of this story is that predictive validity and concurrent validity as defined by a single correlation coefficient are simply not an adequate basis for action. However, if coefficients of correlation are consistent with previous findings and with even a crude theory based on these findings, then there begins to be some basis for generalization, and therefore for justifiable action. In a scientific sense, the only real validity is construct validity.

Some sophisticated reader is likely to point out that the problem here discussed stems from the fact that the prediction study that posed the problem was ill designed in the first place. The investigator should have proceeded by including only those variables that provided some rational basis for believing that they should predict the ability to succeed in law school as it is measured by course grades. If a rational

or are made, indirectly from response conditions. When the counselor is faced with an account of home conditions as told by a student who has serious emotional problems, he may have to make inferences concerning what is objective observation on the part of the student and what is the product of his own distorted perceptions. This difficulty has increased and strengthened the role of the social worker in psychotherapy, for the social worker, among his other activities, can assume the role of direct observer of the conditions that affect the patient's behavior.

From a research point of view, the nature of conditions to which individuals are exposed must be determined directly. Response-inferred conditions such as the clinical psychologist or the psychiatrist is forced to consider cannot possibly form a basis for research that is at all satisfactory. Conditions existing in the classroom should be established independently of what pupils say about them. Insofar as feasible, they should not be described in terms of what a human observer believes them to be, because the process of interpreting events introduces distortion in an unknown direction and by unknown amounts. If mechanical instrumentation can possibly be used for recording relevant conditions, it should be used—not only because of its objectivity, but also because of the permanence of the records that such instrumentation provides.

The Functions of Mechanical Instrumentation

Instrumentation represents an observation technique that is generally much better adapted to experimentation than to field observation occurring under existing conditions over which the experimenter attempts to exert no control. Physical instrumentation, as a result of its very nature, must restrict the recording of observation to a quite narrow channel of events. In the field-observation situation, it is difficult to arrange matters in such a way that this narrow channel can be picked up, recorded, and quantified, without tampering with the situation that it is desired to record.

The first type of instrumentation is likely to serve only the most trivial ends in educational research. While it is possible to measure such factors as the classroom temperature, the illumination on the page of the book or in the classroom as a whole, or the noise level in a particular situation, it is only rarely that some significant relation-

ship is found between such variables and significant aspects of human behavior. Those aspects of the pupil's environment that can be quantified without resorting to instrumentation are likely to represent the most significant aspects of the educational environment. There is no need to resort to instrumentation to determine the number of books in the school library, the number of pupils in the class, the years of training of the teacher, and such items.

Physical instrumentation in research may serve two primary purposes. First, it may simply provide a record of events as they occur. Motion pictures and sound recordings of classroom happenings are of this character. These serve only the purpose of making it possible to reproduce the essential elements of a particular situation again and again, so that the material may be re-evaluated or reassessed in some way by other raters or other observers. The bulk of the material that this involves is always substantial, and the process is costly. Instrumentation that permits reproducibility should be embarked upon only when there has been the most careful planning and where adequate funds are available for the purchase of materials and the employment of personnel.

The second type of instrumentation serves a different purpose. In this case, instrumentation serves more than the purpose of recording events as they occur, for the product is not just a record but a quantification and reduction of events to measures that can be used. For example, it may be possible to determine how much movement the teacher makes around the classroom in a given period of time by equipping him with a pedometer. The pedometer will provide a single over-all score, which indicates the number of steps taken during the period of observation. However, it is rare that instruments which permit the recording of particularly relevant behavior can be attached to a teacher or to pupils in a classroom situation. Instrumentation is much more appropriate to a laboratory situation in which special provision is made both for the appearance of particular aspects of behavior and for their measurement.

It may be pointed out at this time that the use of physical instrumentation necessarily places restrictions on what is to be observed. Most physical instrumentation results in the measurement and recording of only a limited aspect of phenomena. This is true regardless of whether the phenomena are derived from the physical or from the

behavioral sciences. Usually, those selected are believed to be crucial elements of the phenomenon that is being studied, and since few can be measured, it is essential that these be crucial elements and that they be of theoretical significance.

The student of education may not be fully aware of the complicated measurement function of instruments in research. An example may illustrate this function. A relatively simple electronic device may be constructed to record the noise level in a classroom. This device registers the various physical disturbances of the atmosphere that fall within the range of audible frequencies, combines the energy values of these various disturbances, and indicates on a meter some linear or other function of these energy values. The device may be arranged so that it will give an average reading of the noise over, say, a ten-minute period. The instrument can be adapted to provide a numerical reading related to the particular function of noise that it is desired to record. It can thus automatically summate and eliminate the need for numerous readings, which otherwise would have to be summated by hand.

The writer recently observed a striking example of the use of physical instrumentation for recording a limited phenomenon and summarizing the record in a convenient form. The instrument formed part of a device for measuring auditory acuity at various levels of pitch. The person tested listened through an earphone, and each time he heard a buzz he pressed a button. The first series involved a tone at about the pitch of middle C. Tones were given at odd intervals with declining intensity until the point was reached where the tone was inaudible. Then a louder tone of a higher pitch appeared and was presented also at odd intervals in declining intensity. This procedure continued with series of higher and higher pitch. The results of the subject's performance on this test were summarized on a punched card that showed the performance of the individual on the test and the degree of hearing loss, if any, at each pitch level.

Apparatus in Educational Research

Many theses and dissertations in the field of education have had to be abandoned because of apparatus problems, and some have been extended years beyond the expected date of completion for the same reason. It is therefore appropriate that a few comments be made

here in order to steer the student away from some of the deeper pitfalls in the use of apparatus.

First, it is perhaps worth pointing out that the author has known a number of graduate students who insisted on developing research projects involving apparatus that took years to build. Some of these students failed to reach the point of obtaining a doctoral degree simply because they never completed the apparatus. If the student doubts that he can build the required equipment within the space of a few months, he usually should abandon his dissertation problem and find another. Students differ greatly one from another in their ability to build apparatus, and only the student himself can judge his capacity for building experimental equipment. One suspects that the person who has high competence in this respect is also the individual who can show the greatest ingenuity in developing simple designs. Many apparatus troubles result from a failure to simplify apparatus to the point where it will achieve the desired purpose with a minimum number of working parts.

Second, in the planning of apparatus, it is most desirable to incorporate working units that are already available, such as slide-projectors, camera shutters, chronoscopes, and amplifiers. Much equipment that is available around a college can be adapted for experimental purposes.

Third, it is of the utmost importance that equipment be such that it is relatively free from malfunction. Apparatus that breaks down during the course of data collection wastes one of the assets that the experimenter has to conserve most carefully, namely the time given him by his experimental subjects. Freedom from malfunction is partly a function of complexity and partly a function of good design and the use of appropriate materials. Do not, for example, build apparatus in which moving parts are made of wood. These are always unsatisfactory. If metal parts cannot be made, then try using plastics, which will not shrink and expand with changes in the humidity of the atmosphere, as will wooden parts. Plastics such as Lucite are easily bent into desired shapes when warm and can be filed and machined when cold. Electrical contacts are always a serious source of malfunction, and for this reason all connections should be soldered. Where switches are required, it is important that good-quality products be used, and the experimenter should never use the home-made variety.

Particularly useful are modern microswitches and relays. which can be obtained in good quality at a cheap price. Small mercury switches are also cheap and are most satisfactory where small electrical loads are to be carried.

The building of complicated electrical equipment should be limited. It is rarely desirable for the student to work out his own electronic circuits, since those already available are likely to represent the limit of what can be accomplished at the present time. This is particularly true of DC amplifiers, which provide enormous amplification but are also unstable unless they have rather complicated additional balancing circuits.

The reader should perhaps be warned that the cost of having an instrument company build apparatus is usually prohibitive. This statement does not mean that instrument companies charge unreasonable prices, for the actual cost of producing such equipment is high. Not only does it require expensive and elaborate machine tools and hence high overhead charges, but it also requires relatively well-paid craftsmen and highly paid supervisors. In addition, even the smallest part that is to be custom made must be drawn by a skilled draftsman before the task of making it is assigned to the machinest.

Finally, whenever a piece of apparatus is made for conducting an experiment, it is desirable that the entire experimental procedure. once it is started with a particular subject, be completely automatic and not require the intervention of the experimenter. Although there are exceptions, it is by and large most undesirable for the experimenter to have to stop the apparatus from time to time in order to interject some addition to the directions. Thus, if the experiment is conducted by running a sound-recording tape, it is desirable that the full directions to the subject also be recorded on the same tape.

With respect to the use of standard moving picture or sound-recording equipment, a word of caution may be voiced. While the products of such equipment may enable the observer to review events at his leisure, this review is usually a time-consuming and tedious procedure and is rarely worth the effort that it involves.

One ingenious adaptation of a photographic technique may be mentioned. This technique was first developed by Arthur Lumsdaine and was later further perfected by Nicholas Rose. The purpose of these investigators was to record audience responses to moving pic-

tures by means of infrared photography, which permitted the taking of pictures in a darkened room without the audience being aware of the fact. The series of still shots was then analyzed. Figure III shows an example of an audience response recorded by this technique.

Figure III. Infrared photograph of an audience watching a sequence from the film *Life of Riley*. This photograph is the result of a technique for studying audience behavior and is part of an unpublished doctoral dissertation—"A Psychological Study of Motion Picture Audience Behavior," by Nicholas Rose, Ph.D., now Chief Psychologist, Wadsworth Veterans Administration Hospital, Los Angeles 25, California. Photo by courtesy of Universal-International Pictures, and Dr. Rose.

THE DIRECT OBSERVATION OF BEHAVIOR

Up to this point, discussion has not been centered on the direct observation of behavior but on the instrumental recording and quantification of behavior. This procedure circumvents the major difficulties encountered in the direct observation of behavior and the computation of quantitative measures derived from the observed

events. The limitations of instrumentation are such that the educational researcher is inevitably brought up against the problems of observing behavior directly. These must now be considered.

Methods of Observation

Faith in observation would appear to be a cornerstone of teacher education. Some teacher-education programs emphasize the need for sending the teacher in training out into schools to observe at the earliest stages of professional study. Observation, as such, is considered a useful activity. Some, however, would question this procedure, and would say that unless a person knows what to observe the activity may be quite pointless and useless, and that the mere activity of looking and seeing serves little purpose unless certain other conditions have been established previously.

Much the same is true of scientific observation. It would be almost universally agreed that observation is an activity of central importance to the scientist, but it is not just a looking-and-seeing activity. This is a fact that is not always properly appreciated by the person who embarks on his first scientific inquiry. The writer can recall an educator who was starting on one of his first researches, which he decided should be in the general area of teacher effectiveness. He decided that the best way of starting research in this area was to undertake an extensive program of classroom observation. After many hours of this activity, he found that it did not seem to be leading anywhere. Since he felt that his technique of observation might be at fault, he invited some of his graduate students to participate with him in these observation sessions. Much to his surprise, this did not seem to improve matters, and the project was abandoned because it did not appear to be producing results.

The error in this approach lies in the assumption that the mere process of looking at phenomena will reveal what is relevant in them for particular purposes. Conan Doyle in his Sherlock Holmes stories provides an illustration of this fallacious type of outlook. Sherlook Holmes' success is attributed in large measure to his "powers of observation," and it is implied that he is naturally able to see more details in what he observes than are the other characters. This concept is derived from a fallacious psychology, which equates the data of sensory experience with what is perceived. Perception is, in con-

trast, necessarily an interpretive process. In observing a classroom, the sensory data consist of movements of physical objects and vibrations in the atmosphere which are referred to as sounds, but what is perceived is vastly different from this conglomeration of changes in physical energy. What is perceived is an organized, continuing activity, but the concepts and ideas in terms of which the activity is perceived depend upon the experience and training of the observer.

The latter point can be clarified by describing an experience that happened to the author some years ago, when he was invited to participate in classroom observation as a part of a program of research. One of his fellow observers was a clinical psychologist with strong leanings toward the psychoanalytic point of view. The other was an educator with substantial experience in the training of teachers but with only a meager background in current psychological theory. Indeed, the interests of the latter individual were more in the realm of developing specific classroom skills, and were little inclined toward the interpretation of teacher behavior in terms of personality traits and mechanisms that were products of the individual's own background and personal history. These two observers gave entirely different descriptions of what went on in a particular classroom. In his description, the clinical psychologist referred to the extensive oral aggression of the teacher whenever it appeared that the classroom situation was becoming out of control. He also pointed out that such oral aggression (raised voice) was also followed by feelings of guilt, which made her inclined to offer the children various minor rewards, consisting mainly of mild praise. The educator, on the other hand, described the teacher's raised voice merely as disorganized behavior resulting from the fact that she had not acquired genuine facility in using the skills needed for the control of a class of children. What the clinician described as behavior reflecting guilt feelings, the educator described as a return to skillful methods of exercising control over a classroom.

The point here is that any useful description of the tremendous complexities of events in the classroom must be made in terms of a system of interpretation, commonly referred to as a frame of reference. It is necessary for the observer to do more than describe the objectively occurring events, for these are mere movements of conglomerations of matter. What is needed is the abstraction of various

classes of these events, in terms of what are believed to be certain relevant determinants of behavior or to have certain important consequences. The situation is considerably different from that encountered by physicists in observing the moving needle of a galvanometer. The physicists would be able to stay closely with the task of describing strictly what they observed in terms of a simple motion in space. No complex interpretations need to be introduced. The situation to be observed does not involve the vast complexities of a classroom, which, because of the immense number of simultaneously occurring events, can be described only by reducing these events to certain broad but meaningful categories. Writers on research in the behavioral sciences have often compared observation in the physical sciences with that in their own field without observing this essential difference between the two.

The reader may well ask at this point, "Surely would not the observer be performing an observation process comparable to that performed by a physicist when he enters a classroom and notes the frequency of such well-described events as yawning among pupils?" However, even when such well-defined phenomena are involved, judgment is not entirely eliminated. The observer must decide whether a student who opens his mouth just slightly and then closes it is to be considered as yawning. Human activity of even the simplest sort shows a wide range of variation. It is for this reason that the student of behavior must be concerned with the reliability of observation, while the physicist does not have this problem except under very rare circumstances. If human behavior were more stereotyped, the psychologist would not be faced with this problem to the same degree.

What has been said does not lead to the conclusion that observation without clearly defining what has to be observed is always a pointless activity; for the fact is that under certain circumstances it may be useful, particularly in the early stages of an inquiry, but only if it is properly carried out. Suppose that an investigator held the theory that hostile acts on the part of the teacher tended to result in hostile acts on the part of the pupil. He might start by sitting down in his office and listing pupil acts and teacher acts that could be considered as hostile, but his own memory might turn out to be a poor source of materials for this purpose. At this point, he might feel that a better source of information would be the classroom itself. On this basis,

the investigator might well visit several classrooms for the purpose of obtaining lists of behaviors that might be considered symptomatic of hostility. At the same time, he could obtain some estimate of the frequency of each one of these behaviors. Clearly it would not be useful to include in such a list of behaviors those that occur only very rarely, because the investigation might not include a sample of sufficient size to include a single observation in this category.

General classroom observation of the type discussed in the previous paragraph also serves the purpose of indicating the extent to which hostile behaviors are identifiable. While careful reliability studies must be undertaken later, it is important to obtain at an early stage a rough estimate of the extent to which observers can agree on the presence or absence of particular aspects of behavior. Items that are not easily identified may then be removed from the list at an early stage. Sometimes items of behavior that one may expect to be easily recognized do not appear so when an attempt is made to identify them in a classroom situation.

Once this initial stage has been completed, the investigator will have in his possession knowledge of what can and cannot be done in the way of collecting in the classroom data that are relevant to the solution of his problem. It then becomes important to systematize the observation process. This is usually accomplished by preparing a schedule that is to be used by observers in subsequent phases of the investigation. Such a schedule both serves the function of indicating what is to be observed and provides a means of recording the observations.

The Recording of Observations

In research, special problems not met in daily life are encountered in the recording of direct observations of behavior. Consider, for example, the observations that are made by the teacher in conducting her class. These represent a range and variety of facts that are singled out from the vast medley of happenings in the classroom because they appear to have some special significance to the development of the pupil about whom they are made. These facts are observed and recorded for a great many different purposes. They may provide information to be passed on to the school psychologist, the parents, or the teacher of remedial reading, or to be used later in counseling the child.

They may also be used in helping the child learn by giving appropriate assignments or in some other way. Usually, if they are recorded at all, they are set down as brief anecdotes relating the salient elements in the incident. Now while such facts are extremely important from the point of view of running the classroom, they do not constitute the kind of data that the scientist seeks to collect. This fact is often not well appreciated by educators, some of whom have been known to arrive at graduate schools with several crates of such materials and the hope of deriving a doctoral dissertation from them. They are usually embarrassed by the answer to their question, "Here's the data, now what do I do with it?" While the reader at this point will know that the definition of the research problem must generally precede the collection of data, it still may not be clear to him how the data collected by the teacher, and the way in which they are collected, differ from those used by the scientist. Therefore, some expansion in this point is needed.

The first point of contrast has already been made, namely that the teacher collects data for a multiplicity of purposes while the scientist does so for the sole purpose of testing a single central hypothesis.

A second point is that the teacher uses the data as they are collected, while the scientist processes his findings in order that they may be used to answer the questions that have been asked. The scientist cannot use directly the crude data presented by such material as anecdotal records. For the teacher, such records convey the required meaning directly, but for the scientist the meaning is much more indirect. The scientist prefers to handle his data by first reducing them to quantities and then manipulating these quantities through the application of quantitative methods.

Rating as a Method of Reducing Data

Much of the data that is manipulated in educational research must be reduced to quantitative terms by means of a rating procedure. Unless one is dealing with a case in which the rater judges the relative frequency with which an event occurs, the rating procedure involves the evaluation of numerous events, which individually have only small relevance and probably low reliability but which collectively may have value for prediction purposes. If it were possible to design a machine that would identify these events, score them for significance, and

then add up the scores to give a total, and if it were possible for the machine to do this consistently, the results would probably be much superior to those produced by the human rater. The scores or ratings produced by human raters are based on a much less consistent performance than that of our imaginary machine. Actually, the method of reducing data by means of a rating procedure is quite complex, because the rating is usually based not on a well-defined series of events but on a rather vague universe of events, often defined only in the most general terms. If the researcher considers the rater as a rather complicated machine for reducing data, it is evident that in order to have proper control over the measures derived by the procedure, he must know what goes into the machine as well as the nature of the operation performed by it. The fact that there is very little control over the data that our statistical reduction machine is given to use accounts to a great extent for the unsatisfactory nature of ratings. An additional complexity in the use of ratings arises from the fact that two ratings (on two pupils, two teachers, or two whatever else) may not be based on the same data at all. This is not quite as damaging as it sounds, for we may compare the achievement of two pupils even if they did not take the same form of test, and even though the items that they answered were different.

When two supposedly parallel forms of a test have been equated and it has been demonstrated that scores from the two can be used interchangeably, the researcher feels little hesitation in treating scores from one or the other as if they were alike. The equivalence of data derived from two sets of facts must usually be assumed in the treatment of ratings. A rater is employed to derive whatever significance may be possible from the data, and this he does on the basis of judgment, which in turn is based upon experience. An illustration may help to clarify this point. Suppose that one were to rate the children in a class with respect to their cooperativeness with the teacher. Cooperativeness on the part of the pupils might be manifested in a number of ways, such as doing assigned work quietly, attending to the teacher when she is speaking, volunteering to help with chores such as cleaning the blackboard, volunteering information in class discussions, helping other children with assignments, restraining aggressiveness when reprimanded, and perhaps an immense range of additional and varied behaviors. The rater has at his disposal a

sample of these behaviors, but no two children present the same sample, since each shows cooperation or the lack of it in a form that is compatible with his own personality structure. The rater must somehow evaluate the evidence provided by each pupil and make some judgment concerning what that evidence shows concerning cooperativeness. The task is obviously an extremely difficult and complex one, particularly so since the ground rules for the whole operation have not been precisely defined. It is small wonder that the results of many studies that involve ratings produce only a mass of data to which there is little rhyme or reason and from which no useful scientific knowledge is derived.

If control is to be exercised over the rating process so that the product is meaningful, it is necessary to control both the type and quantity of information to be used by the rater and the processing that this information is to undergo. Let us consider the first of these problems.

It is extremely difficult to define for an observer just what is the universe of events to be observed, and, in fact, this is not usually done except in the vaguest terms. For example, a teacher may be asked to rate pupils for their ability to work with other children in small groups. It is probable that the researcher engaged in this enterprise would supply the teacher with a rating scale in which various positions would be described in such terms as "works well with group, seems to add to what the group accomplishes, contributes to the smoothness with which the group operates"; and perhaps at the other end of the scale the statement, "Generally seems to be a source of friction and irritation in a group." Now such a series of statements does very little to orient the rater in the matter of what to observe, but rather it assumes that the rater knows the kinds of observations that are necessary and relevant in order to arrive at the kind of judgment that the scale demands. There is no entirely satisfactory way of remedying this situation. An obvious partial solution is to provide a preface to be read as an orientation to the use of the scale. While such a preface may help to orient the rater on the matter of what he is to observe, it can refer to only a limited sample of the universe of behaviors to be observed, since a long list becomes tedious to read and remember. It may also draw attention to certain specific behaviors, and the rater may easily forget that the behaviors listed are supposed to represent only a sample and not the total universe of behaviors to be observed.

An alternative procedure, which has considerable merit, is to develop a rating scale consisting of many scales, each of which is directed to a fairly specific aspect of the total domain of behavior that is to be observed. If pupils are to be rated on their ability to work in small groups, each pupil might be rated with respect to each one of several aspects of the behavior and perhaps as many as twenty aspects might be listed. When such a procedure has been adopted, it is usually desirable to perform a factor analysis of the ratings to determine whether they can be considered to contribute to a single principal factor. One major advantage of the multiple-rating approach, in addition to the assistance that it gives in defining the domain of behavior to be observed, is that it usually helps to increase the reliability of ratings.

Efforts to Control the Rating Process

Efforts to exercise control over the rating process are familiar, for the common ones are cited in every textbook in educational measurement. The student is undoubtedly familiar with the usual rules, such as:

1. Define several points on each scale with as great precision as possible.
2. Restrict each rating scale to a narrow range of behavior that can be well defined.
3. Change the ends of the scale so that the "good" end is not always at the top or always at the bottom of the scale.
4. Avoid words such as "average" in the middle range of the scale. The rater who does not wish to give too much effort to the rating procedure is likely to class too many as "average."
5. In the directions, indicate the need for honest rating, and, wherever possible, state that a low rating will not have any consequence for the person rated, either direct or indirect.
6. Assure the rater that his anonymity will be safeguarded.

But rules such as these and others, which are useful tips and provide some little help in rating, do not result in the exercise of adequate control over the rating process, at least not the type of control that an experimenter might wish to exercise over the way in which measures are produced. The usual suggestions are not to be disregarded, for they may perhaps convert wholly inadequate procedures into

procedures that, although poor, have enough value to make them usable to a limited degree.

The various attempts to improve the traditional type of rating scale have not produced any instruments that represent a startling improvement over those of several decades ago. It is also doubtful whether any of the more novel approaches to rating have been more successful. One of these, which has been a source of considerable controversy, is the forced-choice approach of the type that has been developed by the Adjutant General's Office. The reader is referred to Guilford's *Psychometric Methods* (1954) for a discussion of this technique, which at this time must be considered of a controversial nature. Unfortunately, it is extremely difficult for a reader to disentangle the merits and demerits of this particular technique, because the enthusiasm of some of its users gives impressions that a more critical appraisal do not justify. Such uncritical appraisal makes it difficult for readers of the literature on forced-choice methodology to determine what has and what has not been accomplished. It is hoped that the reader will review articles on this topic with a much more detached eye than is typical of their writers.

Reliability of Ratings

In theory, if our directions concerning what is to be observed are sufficiently exact, if the observer has been precisely informed concerning the operations to be performed, and if the method of recording the final product of these processes has been well defined, it should be possible for two observers to arrive at closely similar if not identical ratings after observing groups of situations in which there are a range of differences. Interobserver reliability provides some evidence of the extent to which all of these factors have been specified in a satisfactory way. It is possible that good interrater agreement may be achieved even though adequate specifications for the entire procedure have not been provided. For example, teachers may agree on rating pupils for social adjustment even though they cannot provide an adequate definition of what is meant by this characteristic. On the other hand, if all specifications have been accurately made and are capably followed by two observers, it is inevitable that the resulting ratings will agree.

If ratings are to be meaningful, it must be possible to communicate

the rating process so that different individuals can achieve the same results. If the procedure is not communicable, then it is evident that the particular research is not repeatable because of the lack of communicability of the operations that it involves. For this reason, in all studies that involve ratings it is necessary to demonstrate that there is interrater reliability, for lack of such reliability probably indicates lack of communicability of the procedures that the research involves.

There is also considerable value to be achieved in determining the consistency of rating from occasion to occasion. If there is consistency from rater to rater but not from occasion to occasion, it indicates that the phenomenon studied is not a stable one. If teachers were to be rated for some aspect of aggressive behavior shown toward children in the classroom, it is quite probable that raters would agree well among each other concerning the amount of aggression shown on a certain occasion, but the teacher might show little or no consistency in this trait from one occasion to another. Indeed, the amount shown might depend primarily on such factors as the time of day and the presence or absence of petty out-of-school frustrations.

In most rating studies, an effort is made to work with characteristics that have stability over time, but it is quite conceivable that studies might be run in which changeability of the characteristic rated was sought—as, for example, if the researcher were investigating changes in behavior as the new pupil adapted to the school situation. Under such a condition, the researcher would want consistency from rater to rater, but not from occasion to occasion if the occasions were so spaced as to cover a period of time over which changes were hypothesized to occur. Sometimes it may be desired to collect data in such a way that the effects of certain changes on the phenomena to be observed are to be eliminated. Thus, in the hypothetical study of the aggressive behavior of teachers, it might be desired to eliminate variations that occur during the course of a day, and for this reason it might be planned to collect ratings only during the first hour of each day of teaching. By means of an analogous procedure, variations during the course of the week might also be eliminated.

The Interview as an Observational Technique

Up to this point, discussion has mainly centered on the problem of observing classroom and group-activity situations. In such cases

the observer is almost always external to the situation that is being observed. In the interview, on the other hand, it is usual for the observer to be the interviewer and thus to form a part of the total situation within which observations are made. Attempts have been made to introduce observers who are outside of the interview situation, but this is not a usual technique. When the latter technique is used, the interviewer can play a role in which he has been thoroughly drilled and can do so unhampered by recording procedures.

Interviews may vary in the extent to which they are structured. The chief advocates of the unstructured interview have been clinical psychologists, who have used extensively the type of interview in which the conversation is left to wander where it will. The argument has been that, since the causes of particular characteristics of behavior vary from person to person, questions that are appropriate for probing in one case are inappropriate in another. The clinician feels a need to vary his tactics as the situation demands. One consequence of this flexibility is that he is not likely to discover laws that apply to a number of individuals, and indeed that is not what he is looking for.

The researcher, on the other hand, is looking for general laws and has little, if any, interest in the idiosyncrasies that make each patient a unique person. He cannot possibly consider the idea of collecting each item of data under different conditions from every other item. For this reason, in conducting an interview whose data are to be used for general scientific purposes, he attempts to introduce as much uniformity as possible into the procedure. If the interview is to be highly structured, he asks the same series of questions of each person interviewed and does not vary either the order of the questions or the tone of voice in which they are asked. He establishes a uniform procedure that he applies whenever the respondent becomes discursive and wanders too much from answering the question asked. He uses the same introductory remarks and the same way of concluding the interview.

If a structured interview is used, and if it proceeds with an organized list of standardized questions, it may be found desirable to ask questions that are open-ended or those that restrict the possible responses that the interviewee may make. In the former case, the interviewee is expected to *recall* or generate an answer. In the latter case, it is necessary only for the interviewee to recognize the response of his choice. There is some evidence that recognition, at least in a

test situation, produces more information, and more accurate information, than recall. One presumes that this is true also of an interview situation, but there is actually not too much empirical fact to support this view. When the interviewee is free to give any answer to the questions of the interviewer, there is danger that the interviewer may incorrectly record what is said or the gist of what is said. While the amount of distraction undoubtedly varies from situation to situation, there is at least one study by Payne (1949) in which 25 per cent of statements recorded by the interviewer were found to be wrong when they were compared with a recording of the entire interview. Such errors are much less likely to occur when the interviewee indicates to an interviewer which one of a number of statements printed on a card represents his particular choice of response. However, even in the latter case, the procedure for recording the response is not entirely devoid of error. Generally there are also a few persons who refuse to choose any of the responses provided but who prefer to modify one of them before they accept it.

An important aspect of the interview situation is the interviewer himself. For long it was not realized that it is necessary to know something about the characteristics of this individual if the products of the interview are to be evaluated. Those engaged in public-opinion polls have found that interviewers not only tend to select certain types of persons to interview, but the responses of the interviewee are related to the characteristics of the interviewer. Controlling interviewer characteristics is not a matter that can be undertaken easily at the present time. Some characteristics that influence some types of response are known, but others are not. It is also not known to what extent training may result in uniformity of interviewer characteristics.

The novice in research is likely to feel that interviews and methods by which the personal inner life of the individual can be studied offer special promise for yielding knowledge that can be used ultimately for the prediction of behavior. The behavioral sciences started out with this contention, which dominated nineteenth-century psychology. The notion has been perpetuated by psychiatrists, who have consistently advocated the use of individual interviews for selecting persons for special assignments. Yet the rather puzzling fact remains that researches involving inquiry into the inner life of the individual have been extremely unsuccessful. The reasons for this remain quite

obscure, but it is perhaps of some value to the would-be researcher to consider certain sources of difficulty in the methodology of studying the individual, with the hope that this may help the student to avoid them.

First, in any interview in which one person conducts an inquiry into the inner life of another, the situation is much more complex than can be described in terms of an observer and an observed. The situation is more accurately described as involving an observer and a person responding to an observer. The responses are a result of the behavior of the observer and the characteristics of the observed. It is quite possible that relatively minor changes in the behavior of the observer could produce quite pronounced changes in that of the observed. The latter can be clearly seen in the administration of the Rorschach. Any behavior on the part of the Rorschach administrator that indicates that the situation involves threat to the ego of the examinee results in restrained responses. What the observer notes is as much a product of his own behavior as it is a product of the observed's characteristics.

Second, it follows from what has been said that, unless the observer can manifest the same uniform patterns of behavior toward all those who are observed, he introduces a series of quite irrelevant variables into the situation. It is clear that observers are unable to reproduce uniform patterns of behavior when faced with varied situations. Even in actors who in the same play, night after night, face the same situation, there is considerable variation in performance. Many times greater is the variation in the performances of a person interviewing another. This variation is quite beyond the control that an individual can exercise over his behavior.

Third, interviewing procedures are usually based on the assumption that the person interviewed has insight into the causes of his behavior. Clinical psychologists, through experience now covering several decades, have come to the conclusion that insight into the causes of behavior is something rarely achieved, and that even with the extended help of the clinician, it is acquired by dint of long and hard effort. The assumption that it is possible to discover the causes of behavior by means of a short interview is a conception of psychological research that has long been superseded.

Fourth, there are difficulties in quantifying the data provided by the interview. Often the data are such that they do not lend themselves

to quantification. Rarely is it possible to quantify by enumeration, as when the scientist counts the number of words that refer to a given content category. The best that can be done is to rate certain characteristics of the interviewee's behavior.

Fifth, in individual methods of appraisal, the psychologist is often looking for the laws of the behavior of the individual. Insofar as these can be discovered, they yield qualitative statements that summarize trends in past behavior. but these trends may not be related to the making of the desired type of prediction. In contrast, objective methods of appraisal are always designed to measure variables that are empirically or rationally related to the variables it is desired to predict. The material derived from an interview may have only the most remote relevance.

Sixth, a person who is being studied by another may not be willing to give himself away. There is a real difference between the behavior of a person who visits a clinical psychologist in order to seek help and a person who is not motivated in this way to lay bare his innermost thoughts. In the latter case, there is a certain defensiveness about the individual's performance, and an unwillingness to reveal what is in his mind. Indeed, there are some who refuse to answer the simple questions asked by public-opinion pollsters because they say that this would be an infringement of their privacy. The same difficulty arises in the administration of certain types of tests and particularly instruments of the projective type. The patient at the clinic may be expected to give a much richer range of responses on the Rorschach than the one who is taking the test as a part of some research study.

Finally, what has been said here does not mean that it is not possible to study individual cases over a period of time, for this can be done by many methods that permit the use of objective measuring devices.

In interview situations, unless it is otherwise desired, the greatest caution must be exercised lest the questions themselves convert the situation into a threatening one. In conducting a study of changes in attitudes from age twenty to age fifty, it would be most unwise to ask the fifty-year-old group how they voted *when they were young*. This would imply immediately that the fifty-year-olds were regarded as an aged group, an implication that might be quite unacceptable to them. Often it is possible to reduce the potential threatening effect of ques-

tions by implying that the phenomenon is a common one. If one asks a married person how long it was after he was married that he first contemplated separation or divorce, he is more likely to admit that such ideas have entered his mind than if asked whether he had ever contemplated separation or divorce. The first of these questions carries with it the implication that thoughts concerning divorce or separation are common, or even that they occur in the mind of every married person. The second question implies that such thoughts occur only in certain married persons. Sometimes the element of threat can be reduced by a mere change of terms. It is probable that married couples will be much more willing to talk about sources of friction in their marital relationships that they would be to discuss the things about which they quarrel. The latter is likely to arouse feelings of guilt to a much greater extent than the former.

At this point, the reader may well ask, "Why not just pass out a questionnaire? Why go to all the trouble of conducting an interview when both the questions and the responses are standardized?" The answers are to a considerable extent a matter of opinion, but there would nevertheless be substantial consensus among experts.

First, and this appears to be well established, it seems possible to obtain a much higher percentage of respondents with an interview than when questionnaires are handed out. Interviewers commonly report less than 5 per cent refusals to answer questions, while returns from mailed questionnaires rarely exceed 40 per cent. There is also a tendency for those who answer questionnaires to omit the answers to some questions, through either forgetfulness or a distaste for facing the particular issue. Incompleteness of returns is rarely found when data are collected by interview.

Second, the interviewer is able to answer questions concerning the purpose of the interview, and the interviewee may be put at ease in a way that is not possible with questionnaire techniques. He is thus able to build up a feeling of confidence that makes for both co-operation and truthfulness.

Third, questionnaires present difficulties to persons of limited literacy, and the respondents to a mailed questionnaire study are likely to represent an undue proportion of the more literate public. In addition, persons who read with difficulty may not exercise the care that they should in finding and selecting the right answer, while the interviewer can take care of all such mechanical details.

Fourth, an interviewer can conduct an interview at a proper speed, while questionnaires are often filled in hurriedly. The writer can recall having to fill in a questionnaire late at night in order that it be available for collection on the following morning. True, the questionnaire had lain around the house for two weeks, but somehow time had not been found to answer the questions. This kind of problem can be avoided if an interview is used.

On the other side of the picture, it may be said that it is often much more feasible to present in questionnaire form long and extended lists of questions which it would be very tedious and expensive to present orally. Such questionnaires when administered to highly literate groups may yield large quantities of information that could not easily be obtained by the oral interview.

The Situation in Which Observations Are Made

The observational techniques selected for use depend upon the nature of the situation in which observations are made. If the assumption is made—and it is generally considered a reasonable one—that behavior is a function of both the situation and the personal characteristics of the individual, then it follows that it is necessary to arrange situations in such a manner that the behavior to be observed will emerge. This is generally recognized in experimental psychology, but rarely in educational research. Many studies have been conducted in which children have been observed both at work and at play, but in which no attempt at all has been made to control conditions existing at the time when the observations were made. Such a procedure is similar to administering a different test of achievement to each one of a number of children and then comparing the test score of one child with that of another. In the case of tests, comparability of the instruments used is an essential condition for comparing the scores, and this is rarely forgotten. However, where the situation used for measurement is not a test but a free-play or work situation, this necessary condition is often forgotten. Again, when interviews are used as situations in which observations are to be made, rarely is it found that there is any comparability among them.

There are certain aspects of all situations used as bases for observational data that have to be controlled if meaningful data are to be collected. First, there is the orientation to the situation as such. Usually this is a verbal orientation, and great care must be exercised

in this process. It is not just when social behavior is being observed that this orientation procedure is of crucial importance. Even when a test is to be administered, the orientation procedure is a matter that may be of major importance in determining the score. It is now well known that, in the case of the Rorschach, the responses depend to a very great extent on the circumstances surrounding the administration. If the orientation is such as to imply that the scores might be used in some way that vitally affects the person's future, it is likely that the performance on the test will be characterized by reticence or reserve. Whether the person knows it or not, he is likely to withhold information and to be much less free in giving responses when there is some threat involved in the situation than when no threat is involved. Much the same is true in any other situation in which persons are observed or tested, and it should be noted that the threat involved may be implied rather than stated. If a person, due to the serious attitude of the experimenter, is made to feel that whatever is being answered represents a characteristic of great importance to success in life, he will behave differently than if he feels that the procedure is just an experiment and of no consequence to him personally. The implications for the person under study must be clearly spelled out.

Second, if stooges are used as a part of the situation to which persons are to be exposed, it is important that they be well trained so that they respond in the same way to all. Those who are to function as a part of an experimental situation must learn not to respond differently to different personalities. Learning in this respect is probably only partial, since most persons are not aware of the extent to which they are responding to others.

Third, it is important to develop safeguards so that information about the situation is not transmitted from those who have been exposed to those who are still to be exposed. Sometimes the mere separation of subjects from others who have been exposed will suffice. Sometimes other precautions are necessary, such as the selection of persons from different classes or from different schools. It may help if subjects are asked not to divulge information, but even when this is done, there is likely to be some leaking of information.

Finally, it is hardly necessary to point out that some of the more obvious factors affecting behavior, such as time of day, should be carefully controlled, or designs that permit the separation of the variance attributable to these factors should be used.

Role-Playing as an Observational Technique

An interesting extension of the interview technique, in which there has been much interest in recent years, is found in role-playing. An example is given in an unpublished study in which it was desired to explore the personalities of prospective teachers by means of this technique. For this purpose a number of situations were developed. In one of these, one of the participants was given a rather detailed account of the behavior of a pupil in a class that the student was theoretically teaching. The student was given fifteen minutes in which to study the material, and was told that at the end of that period she would have to interview the mother of the pupil in order to discuss with her the problems of the child. Another student playing the role of the mother was instructed to take a very defensive attitude toward the child and a hostile attitude toward the teacher, placing the blame for the child's behavior on the way in which the school was managed. The scene was then spontaneously enacted by the two participants.

The reader can see that there are extensive possibilities for personality measurement and assessment in such a situation. The behavior of the persons involved can be rated for various characteristics, and there are also possibilities for using check lists for enumerating the frequency of occurrence of specific aspects of behavior. While the technique is still in the exploratory stages of development, promise is offered by the fact that persons placed in role-playing situations become deeply involved emotionally. Many play their parts as if they were completely identified with the character portrayed. The technique cannot yet be considered as a ready-made usable product, but it is an interesting invention.

Role-playing also has potentialities as a training technique. It offers promise as a direct means of teaching persons to handle rather complex social situations. Here again there is a fine field for educational research.

The Usefulness of the Observational Techniques Reviewed

When educational psychology began to develop over half of a century ago, it was hoped that relatively simple techniques for observing and recording aspects of behavior would yield information of great value to the educator. This hope has not been fulfilled. Simple

rating procedures have not proved themselves to be a satisfactory way of assessing relatively enduring personality traits. Nevertheless, researchers will probably continue for many years to come to conduct studies in which ratings are used, because it is believed that the small amount of information provided is better than no information at all. Also, at the present time there does not appear to be anything in sight that might be considered an improved substitute.

In recent years it has been hoped that a substitute for rating might be found in measures derived by placing persons in standardized situations and measuring aspects of their performance in these situations. This approach, which has commonly been referred to as an assessment procedure, has failed to yield results that can be considered in any way superior to those obtained from traditional procedures. Such procedures are based upon the assumption that personality traits that appear in one situation will appear in other and different situations also. This simply does not seem to be the case. A person who is aggressive in one situation is withdrawn in another. A person who is happy when he is in one place is unhappy in another. A theory on which observation is based must ultimately take into account the variability of behavior from one situation to another.

Summary

1. The scientist uses the term "observation" in a rather different sense from that ordinarily used. The scientist refers to an item of data as an observation.

2. There are two distinct classes of observations that are made in the behavioral sciences. They may pertain to the characteristics of the environment or to the responses of living organisms to that environment.

3. Instruments may be introduced into the observation process for various purposes. They may improve the precision of the resulting observations. They may eliminate the need for human observers over long periods of time. They may summarize observations that would otherwise be bulky and cumbersome to handle.

4. As far as it is feasible, any instruments or devices used as a part of a research should consist of standard parts that are easily maintained and highly reliable in operation.

5. The undeveloped state of the art of educational research is such that instruments have only the most limited value. Most of the data that such research involves must be collected by other procedures, and direct observation is probably the most widely used of these.

6. Observation is a procedure for classifying and recording events according to some plan or schema. The latter is commonly referred to as the frame of reference of the observer. The observer should approach a situation with a clear idea about what is to be observed and with some rudimentary theory about the significance of his observations.

7. The classroom data collected by the teacher usually differ in many respects from those collected by the scientist. One marked difference is that the teacher uses his data directly, while the scientist uses them only after they have been processed in some way.

8. Rating is a way by which numerous events that are observed are somehow summarized and combined. Control over the rating process requires that control be exercised over both the nature of the information that is used and the way in which it is used. The unsatisfactory nature of most ratings is a reflection of the difficulties involved in introducing such controls.

9. Any rating procedure developed should be such that it can be communicated to others.

10. The interview is a commonly used technique for obtaining observations concerning individual behavior. It presents a highly complex situation from which it is difficult to obtain data that can be reproduced in other studies.

11. While interviewer characteristics are important variables in the situation, they are difficult to identify and control. It is even difficult to train interviewers to make accurate records of what happens.

12. The interview must be considered as a complex social situation in which the interviewer and the interviewee are making continuous adjustments to the responses of one another.

13. In planning interviews, caution should be observed in making assumptions about the extent to which the interviewee has insight into his behavior.

14. The interview has certain advantages over the questionnaire, such as a higher percentage of respondents, superior cooperation, and a lack of dependence on the ability to read.

15. Attempts have been made to place persons in situations that represent reproductions of real-life situations, in order to observe their behavior.

Some Problems for the Student

1. In a certain teacher-training department, the supervisor of each student teacher is required to observe him conduct several hours of classes and then to submit a report on what he has noted. What are some of the factors in the supervisor's background that might result in a

tendency for the observation of student behavior to be selective? Suggest methods that could be adopted to insure that the observations submitted by different supervisors would be comparable, at least to some extent.

2. Prepare a rating scale or other instrument to be used in the process of assessing the amount of anxiety displayed by persons in a face-to-face counseling interview.

3. When they know they are being observed, children and teachers behave in a manner different from that when the observer is absent. Suggest several ways that might be used to overcome or partially overcome this difficulty.

4. List some personal characteristics that it might be feasible to rate accurately in an interview situation. List some characteristics that probably could not be accurately rated in the same situation.

Observation: More Complex Procedures and Indirect Approaches | 9

In the previous chapter the reader was presented with a discussion of the problems of observing and recording the characteristics of individual behavior. However, most behavior that is likely to be observed as a part of an educational inquiry is likely to be seen under rather complex social conditions, such as those that occur in a classroom. While most rating schedules have been developed for use in situations where the person who is performing the rating is highly familiar with the person rated, observation in the classroom situation presents an entirely different circumstance. The observer in the classroom has probably never before seen the pupils he is observing, and he will probably never see them again. For this reason alone, the type of rating schedule that has been considered is unsuitable as an instrument for recording aspects of behavior seen in most classroom observation situations.

The types of instruments that are now to be considered have been developed mainly for the purpose of helping the observer organize his work when he is faced with a novel social situation such as a

classroom. The schedule is a means of controlling the observer's behavior so that the observations he makes are not based on his personal whims but are those that serve the purposes of the research that is being undertaken.

Classroom Observation Schedules—Problems in Their Development

The reader can turn to a textbook on educational measurement in order to acquire information about the development of rating scales, but he will not find information about the development of observation schedules so readily available.

The development of observation schedules to be used in the recording of events in the classroom is a matter that requires considerable technical knowledge. In order to familiarize the reader with some of the problems that this involves, the development of an observation schedule as it occurred in one study is presented here. The study selected is that by Morsh (1955), which is also discussed in other parts of this book. Examples of other observation schedules are then presented.

In his study, Morsh was concerned with determining the relationship between instructor's and student's behavior and the amount students learn. The study was conducted in technical courses given in Air Force schools. In this research, it was decided to study behavior at a level that involved fairly small segments of what would commonly be called specific behavior. Observers were sent to classrooms in order to obtain lists of what seemed to be relevant teacher behaviors that could be postulated to effect the course of student learning, and also student behaviors that could be hypothesized to be symptomatic of efficient and inefficient learning. On the basis of these preliminary observations, a list of 160 behaviors was prepared. But it was soon found that observers who used it as a check list could neither keep all the items in mind nor observe on such an extended front. This difficulty resulted in the reduction of the check list to 80 items. While the writer does not have data on which to base any criticism of an 80-item check list, it does seem to him that even the abbreviated list included too many items for practical purposes. The alternative is perhaps equally unsatisfactory, for it involves either using a few items that may cover only a limited range of behaviors, or using broader regions of behavior that are likely to be rather vaguely defined.

At this stage of the inquiry, certain observational difficulties soon became apparent. The first of these was mechanical. It simply did not seem feasible to provide observers with a check list that extended beyond a single page. Searching through more than one page in order to record an entry was not a feasible task, particularly when, as in any classroom, much was happening and several different but relevant behaviors occurred concurrently. Second, there was little point in the inclusion in a check list of those items that happened so rarely that they were not likely to be noted by the typical observer in the study. Third, the observer had only limited time in which to record his information and thus items that required some reflection before the decision to record was made should not have been included unless they provided information essential for the study. (There is, of course, no clear-cut line to be drawn between items that involve judgment and items that do not. Rather it is a matter of degree. In many areas of observation, if judgment items are discarded, little of value remains. A simple and familiar example of this is the typical English theme or composition. Systems of judging such compositions that are limited to objectively observable items, such as spelling, capitalization, agreement of subject and verb, etc., have been found to measure only the most trivial aspects of teaching in English.) Fourth, it was considered desirable that the lists be usable by an observer who had had only very limited amounts of training. Fifth, it appeared to be important to make the lists short enough to be memorized by the observer. (The writer feels that this is an extremely important property of all well-built observation schedules.) Sixth, only items were to be retained that had some logical relevance to the learning process as it occurred in the classroom. (This is a point that has been stressed all the way through this book and is merely the application of one aspect of acceptable scientific methodology.)

As a result of these practical considerations, still further reductions were made in the check lists. That for recording observations pertaining to the teacher's nonverbal behavior consisted of a list of thirty-five items. Another list of thirty-three items pertained to the instructor's verbal behavior. A third list of twenty-five items was used for recording observations of student behavior.

Three observers were given preliminary training and were then assigned the task of collecting data in fifteen-minute observation periods in thirty classes. On the basis of these data, the reliability

(consistency) of rating was determined for each item in the check list. Through this computation most of the items were found to have adequate interobserver reliability. Those that did not were almost without exception items that occurred very infrequently. Examples of such items (sad to say) from the instructor's verbal behavior were "praises student, praises class, admits mistake." From the list of nonverbal behavior, infrequent items included such matters as "uses blackboard for key term, checks time, ignores student answer." All of the items except one in the student behavior check list showed satisfactory interobserver reliability.

Now Morsh was aware of the fact that raters might well agree among one another and in this sense the measure might have reliability, but this would not necessarily mean that the measure would have consistency from occasion to occasion. Indeed, from the evidence presented up to this point, it is quite possible that the incidents that occurred during the particular fifteen-minute periods were not typical.

In order to determine whether the observations during the particular fifteen-minute periods of observation could be used to generalize about behavior in other periods, a further study was made. In this subsequent study, each observer visited a large number of instructors for six fifteen-minute periods. These were combined into three half-hour periods for the purpose of determining reliability of observation from occasion to occasion. As a result of this procedure, it became necessary to eliminate certain additional items from the check list because they did not show sufficient consistency from occasion to occasion. As a result of this elimination procedure, there remained thirteen items on the instructor verbal behavior form, twelve items on the nonverbal behavior form, and ten items on the student behavior check list. The lists were then set up for six five-minute periods of observation to provide a total period of thirty minutes of observation on each form. What were some of the items that were retained as reliable after such an elaborate process of elimination? In the case of the student behavior check list, the following items were retained:

Talks
Answers question
Asks question
Looks around
Doodles

Ignores instructor
Slumps
Yawns, stretches
Sleeps or dozes

While the extent to which the student manifests such behavior can be considered as evidence of conditions unfavorable to learning, the same cannot be said of the teacher behavior items. The final teacher behavior check list included an excess of items such as "stands behind desk, stands at board, stands at side, moves, leans on desk, sits at desk." Undoubtedly such items tended to be retained by the procedure because they could be observed reliably. On the other hand, there were also retained some items that it seemed reasonable to suppose were highly related to the teaching process. Examples of these were "ignores student with hand up, smiles, demonstrates at board." The data selected demonstrate the tendency for highly reliable items of behavior that refer to gross bodily positions to be retained, while the more difficult to observe and more subtle aspects of teacher behavior, are rejected in any tryout or procedure for the screening of items.

There is much to be learned from this. The principal lesson is that aspects of teacher behavior that are likely to be related to learning effectiveness because they are difficult to rate will probably have to be rated or assessed through instruments that the investigator has prepared after long and painful effort. This may mean that the researcher will have to experiment with the rating (or checking) of numerous different aspects of behavior until those that can be reliably rated or identified are identified. A satisfactory instrument for use in observing teacher behavior can be prepared only after prolonged effort.

While the previous discussion was presented in order to illustrate the problems of preparing schedules and check lists for recording observations, it would be unfair to the reader not to present some of the consequences of this study. As might be expected, the items in the teacher behavior check list showed little relationship with the extent to which the students showed gains in achievement (corrected in terms of both their initial knowledge and relative level of ability). The difficulty of recording relevant aspects of teacher behavior are indicated by these results. The only alternative explanation would

be that teacher behavior does not provide important variables in the teaching process. The latter just does not seem to be an acceptable hypothesis in terms of what is known about conditions affecting human learning.

In contrast, the behavior of the students as recorded on the check sheet was indicative of the extent to which learning was taking place. This is, of course, entirely reasonable. Students who are yawning, dozing, or sleeping cannot be expected to be learning with any degree of effectiveness. Morsh draws the interesting conclusion that if the supervisor visiting the classroom wishes to make an assessment of the amount of subject-matter information that is being acquired by the student, then he might do well to observe what the students are doing. Observable student behavior may provide more valid evidence of teacher effectiveness with respect to certain goals than can the information derived from the observation of teacher behavior. When we learn what aspects of teacher behavior to observe in this connection, this statement perhaps may need modification.

The Morsh study has been presented here because it is one of the better-designed efforts at obtaining observational data from a class situation. It illustrates some of the precautions that must be taken and the care that must be exercised. It also reflects the fact that the screening process may eliminate from consideration all but the most trivial of observations. When this happens, the researcher should not be content to pursue his inquiry with the reliable but trivial. He has a choice of making another attempt to observe relevant events or of dropping the investigation.

Many of the difficulties that are presented in the Morsh study stem from the fact that we are only beginning to learn useful techniques for classroom observation. The complexity of the phenomena are such that any quantification procedure involves difficulties in abstracting that segment of behavior to be observed, and further difficulties in the quantification of aspects of that segment. New ways of doing this will be constantly explored during the years to come, and it is the writer's belief that methods far superior to any available at the present time will be developed. How far it will be possible to replace human observation and judgment by physical instruments that record and quantify is a matter about which there can be only speculation at the present time.

The difficulties that Morsh tried to avoid led him directly into other difficulties arising out of the specificity of the phenomena recorded. Others have recognized this difficulty and have attempted to devise methods that circumvent it. All of these are based on the concept that a whole category of different behaviors may be used to measure a dimension of teacher or pupil behavior, much as the items in a test may be used to measure a relatively homogeneous and single dimension. There are, of course, innumerable different ways of doing this, because the classroom presents a vast range of phenomena that can be observed, and these can be classified in a great variety of ways. On this account, most of the techniques that have been proposed restrict observation to a limited phase of behavior. For example, a technique proposed by Withall (1949) confines the domain of observation to statements made by the teacher. This is done on the assumption that most of the important interactions that occur in the classroom are undertaken through the verbal medium, and hence a study of verbal interactions will reveal most of the important events occurring there. These statements are sampled according to some plan that provides that they be as representative as possible of the statements made by the teacher in the classroom. Once they have been recorded—and this is done at the time of observation in the classroom—they are available for later classification, which can be undertaken by persons who were not involved in the original recording process. One may then determine the reliability of the classification procedure. The classification of verbal behavior proposed by Withall uses the following categories:

1. Learner-supportive statements that have the intent of reassuring or commending the pupil.
2. Acceptant and clarifying statements having an intent to convey to the pupil the feeling that he was understood and help him eliminate his ideas and feelings.
3. Problem-structuring statements or questions which proffer information or raise questions about the problem in an objective manner, with intent to facilitate learner's problem-solving.
4. Neutral statements which comprise polite formalities, administrative comments, verbatim repetitions of something that has already been said. No intent inferable.

5. Direct or exhortative statements. with intent to have pupil follow a recommended course of action.
6. Reproving or deprecating remarks intended to deter pupil from continued indulgence in present "unacceptable" behavior.
7. Teacher self-supporting remarks intended to sustain or justify the teacher's position or course of action.

These statements are quoted from Withall since his wording is so evidently carefully selected.

Withall's technique is of particular importance because it is based on a theory concerning the determinants of certain classes of events in the classroom. He hypothesized that the teacher's behavior is the most important single factor in creating classroom climate and that the teacher's verbal behavior is a representative sample of his total behavior.

Subsequent work has confirmed the view, expressed by Withall and substantiated by his data. that the statements of teachers can be categorized with reliability. At the time of writing, there is little evidence to show that the measures derived from this categorizing process represent conditions important to the learning process. but on the other hand the categories offer promise. They appear to be good candidates for this purpose since they represent conditions that are commonly postulated to be important factors related to learning. and particularly those related to reward and feedback. In this respect. the study of Withall is in many ways ahead of that previously discussed. for it deals with broad categories of behavior that, at least in other contexts. have been shown to be related to the learning process.

Some of the advantages of the Withall technique over the more familiar observational and rating techniques that have already been discussed may be pointed out. It permits the summation of scores derived from a great number of behaviors. In this way. it is possible to build up reliable scores from a great variety of classroom events. The work and thought of categorizing behavior. which may occupy so much of the time of an observer in the classroom and may detract from his observations. is undertaken at leisure after the observations are made. The technique also recognizes the overwhelmingly important role played by verbal behavior in structuring events in the

classroom, a fact that has been commonly overlooked in most observational techniques.

Some Examples of Observation Schedules

In contrast to Withall's technique, the observational method proposed by Cornell, Lindvall, and Saupe (1953) does not restrict itself to the observation and recording of verbal behavior but covers a much wider range of events. The purpose is broader, though possibly vaguer, and is summarized in the following words: "Our chief concern is with measures or descriptions of 'what students do in school, what teachers do in school,' or in short, just what a school is in terms of the total learning environment it provides the students." Some further restriction is placed on the domain to be observed by the statement that this is to be restricted to events within the classroom situation. It is clear, then, that these workers are concerned with the development of a series of dimensions of the educational environment of the school child, and the hope is that the dimensions will differentiate between school systems. In actual fact, these workers are concerned with more than differentiating school systems. Mere differentiation is not a sound basis for selecting a characteristic for measurement. If the walls of one school consisted of plaster and those of another of plywood, this would not be considered adequate as a basis for selecting this particular characteristic. Any characteristic selected for differentiation must be one that has relevance to the educational process itself. It should be selected not only because it differentiates but also because it is related to some important aspect of the learning process as it takes place in schools. This is implied in the statement that "the most direct impact of the school upon the child is in the classroom." This is a somewhat vague recognition of what we have been saying here. It does not provide a theory of behavior that is a sufficient basis for selecting those differentiating variables that are relevant to the educational process from those that are irrelevant.

The following dimensions for measuring classroom differences are proposed in this system:

1. Differentiation. This dimension defines the extent to which classrooms provide for individual differences both in intellectual and in nonintellectual characteristics. While there is no universally accepted method of providing for individual differences, a number of

common procedures that may be scored along a common dimension are adopted for this purpose.

2. *Social organization.* While this is referred to as a dimension, it seems to the present writer to be a complex consisting of a number of dimensions such as single-group versus multigroup organizations, pupil leadership versus teacher leadership, and group-leader interaction versus no interaction.

3. *Initiative.* This dimension reflects the extent to which the pupil directs the learning process or is directed by adults. The amount of pupil initiative manifested is a dimension in which schools show striking differences.

4. *Content.* This again must be considered as a complex rather than a single dimension. It includes observation of such varied phenomena as one textbook versus several textbooks, the extent to which sources other than textbooks are used, and the extent to which the curriculum is organized around areas of pupil interest or around areas designated by the teacher.

5. *Variety.* This dimension is used to measure the extent to which there is variation in classroom procedure. It can be measured, for example, by the number of different activities that are observed to occur during a given period in the classroom.

6. *Competence.* This category represents a dimension in which the teacher is appraised for the extent to which he is inactive or to which he shows positive behaviors that can be reasonably assumed to be related to learning. It is not competence in the sense in which a teacher is said to be effective or ineffective in the achievement of certain goals, but rather is it a difference between playing a passive or innocuous role and playing a positive role in the classroom. The word *competence* is not too well chosen.

7. *Climate-teacher.* This dimension refers to the extent to which the teacher behaves in a way consistent with the development of good human relations. It deals with interpersonal relations and matters of warmth and friendliness.

8. *Climate-pupil.* This dimension refers to the extent to which the pupils either respond positively to the classroom situation or tend to be restless and inattentive. It does not refer to what is ordinarily referred to as "classroom atmosphere," since this characteristically

includes reference to the degree to which there is tension in the situation.

Regardless of the merits of the various dimensions that have been discussed, there are certain difficulties involved in the use of an observational technique of this kind. The person who applies it uses a sheet entitled "Classroom Observation Schedule" for recording his entries. On this he records his observations by five-minute intervals. However, the observations cannot be recorded directly because the schedule calls for a coded record of what is observed. This is accomplished by means of a "classroom observation code digest." The latter is a two-page affair that permits the numerical coding of what is observed. Only the number derived from the code is recorded on the observation schedule. In addition, there is a sixteen-page manual that describes in detail how this process is to be accomplished; it includes a great deal of detail, and also special instructions for handling various aspects of the coding and for dealing with exceptional cases.

The entire procedure quite clearly calls for a great deal of practice on the part of the person using it. If data were to be collected in a specified group of classes, the observers to be used would have to be given practice on another group of classes. It is doubtful whether a few hours would provide anywhere near the amount of experience needed to obtain real facility with the system or the minimum facility necessary to provide records with good inter-rater agreement. The defect in the system stems from the tremendous burden of work that it requires the observer to do in the recording of observations. This is done in the interests of providing accurate data in a form that makes them utilizable for subsequent quantitative analysis. Conceivably, broad over-all judgments of the type that were commonly made in classroom studies of the late 1930's may yield as much information as those that deal with the more modern type of observation schedule.

In order to give the reader a conception of the range of observational devices that have been developed for recording and rating events in the classroom, two more specimens will be considered here. While the examples previously given are appropriate for recording events in classrooms as they are typically managed, the two to be considered were developed for recording behavior in small-group situations. The opportunities for observing such behaviors in educational situations are extensive. The reader should be reminded again that we are

attempting to present only a small sample of the instruments that have been developed. Those that are presented are chosen because they illustrate certain points. The fact that an instrument is mentioned here does not constitute endorsement.

One example is a system developed by Bales (1954) for recording certain aspects of the interaction process as it occurs in small groups. Bales uses the following twelve categories:

1. *Shows solidarity*, raises others' status, gives help, reward.
2. *Shows tension release*, jokes, laughs, shows satisfaction.
3. *Agrees*, shows passive acceptance, understands, concurs, complies.
4. *Gives suggestions*, direction, implying autonomy for others.
5. *Gives opinion*, evaluation, analysis, expresses feeling, wish.
6. *Gives orientation*, information; repeats, clarifies, confirms.
7. *Asks for orientation*, information, repetition, confirmation.
8. *Asks for opinion*, evaluation, analysis, expression of feeling.
9. *Asks for suggestion*, direction, possible ways of action.
10. *Disagrees*, shows passive rejection, formality, withholds help.
11. *Shows tension*, asks for help, withdraws from field.
12. *Shows antagonism*, deflates others' status, defends or asserts self.

While some of these categories may seem a little vague to the reader, it must be said in fairness that Bales also provides an extended description of what is to be included in and excluded from each of them.

A much more detailed system of categories is provided by Carter and his associates. Their system has been developed in order to record certain aspects of group behavior as it occurs in a laboratory situation. In this system of recording observations, behavior is classified under seven major categories, which are:

1. Showing personal feeling
2. Proposing and initiating action
3. Disagreeing and arguing
4. Undertaking leadership roles
5. Undertaking worker roles
6. Performing abortive and nonproductive behavior
7. Miscellaneous

Under each one of these major categories are listed more specific elements, which represent the level at which behavior is actually recorded. The relatively specific elements in each category are as follows:

Shows a personal feeling of:

1. aggressiveness or anger
2. anxiety or insecurity
3. attention or readiness
4. confusion
5. cooperativeness
6. deference
7. dissatisfaction
8. formality or reserve
9. friendliness
10. negativism or rebelliousness
11. satisfaction or accomplishment
12. status

Proposes and initiates action:

21. calls for attention
22. asks for information or facts
23. diagnoses situation
24. asks for expression of feeling or opinion
25. proposes course of action for self
26. proposes course of action for others
27. supports or gives information regarding his proposal
28. defends self or his proposals from attack
29. initiates action toward problem-solving, which is continued or followed
30. supports proposal of another
31. agrees or approves
32. gives information
33. gets insight
34. general discussion concerning task
35. expression of opinion

Disagrees and argues (with a somewhat negative connotation):

40. disagrees or is skeptical
41. argues with others

42. vigorously argues with others
43. deflates others
44. gives bald comments or prohibits (in disagreeable fashion)

Leader roles in carrying out action:

50. gives information on how to carry out action
51. praises, commends, rewards
52. expresses desire that something be done
53. asks for assistance for others
54. asks for assistance for self
55. integrates group behavior

Follower and worker roles in carrying out action:

60. follows suggestions or directions
61. offers to help or helps
62. imitates others
63. asks for permission
64. collaborates with others
65. answers questions
66. performance of simple work unit (group-oriented)
67. performance of simple work unit (an independent effort)
68. passively helps

Abortive or nonproductive behavior:

70. initiates action that is not followed or continued
71. verbal interplay without outcome
72. listens, but doesn't express self or participate

Miscellaneous:

90. stands around doing nothing
91. engages in "out-of-field" activity
92. engages in incidental conversation while working

Some attempts have been made to develop mechanical devices to facilitate the recording of human observations. One of these, by Bales and Gerbrands (1948), has received considerable comment in the literature. This device provides for a set of categories of behavior into which observations are to be classified. They are listed one below the other as they might be on this page. The list is placed on top of a mechanical device that moves a strip of paper tape past the right-hand margin of the list in a right-to-left direction. The observer marks on

the tape entries opposite the appropriate category whenever events occur that have to be recorded. Once an entry has been made on the tape it quickly moves behind the list. The machine makes a mark on the tape every minute, so that the observations may be analyzed for each one of a series of time periods. In addition, a light may be made to flash on at regular intervals in order to remind the experimenter to undertake particular observations.

The interaction recorder is a convenient device that can facilitate the recording of data and provide a permanent record of the time sequence of events. In the latter respect, the record does not have the clumsiness and voluminousness of moving picture records or sound tapes. The device also provides material on which the observations of two or more observers may be compared event by event, and thus reasons for discrepancies can be discovered. Two records may be compared to obtain rapid estimates of interobserver reliability.

Distortion in Observation

Instrumentation is introduced into scientific inquiry because of the limitations of the human observer. In physics, the rapidity of many phenomena requires instrumentation in order that highly transitory events that could not be observed can be reduced to events that can be observed by the experimenter. In this field, instrumentation not only reduces nonobservable phenomena to observables but also permits the measurement of observables. In the behavioral sciences, instrumentation serves all these ends and also a very important additional purpose. That purpose is to prevent the observer from introducing into his narration events that never occurred.

Real differences exist in the ability of individuals to report observations without bias. McPherson (1954), who studied this problem of the differences between high distorters and low distorters, came to the conclusion that these two groups were distinguished by the following characteristics:

Low Distorters	*High Distorters*
1. Able to integrate the content with their own ideas and the ideas of others.	1. Parrot the content and gives little evidence of understanding the material they are trying to present.

Low Distorters	*High Distorters*
2. Tend to relate the ideas introduced by group members to the content and to each other. Will clarify errors and misinterpretations made by group members.	2. Tend to avoid questions about content and accept misinterpretations without attempting to correct them.
3. Demonstrate a freedom with material by using it as a basis for introducing relevant ideas that serve to broaden or extend the range of ideas.	3. Escape the material by using it as a springboard for introducing highly personal experiences that do not forward an understanding of the content.
4. Restate comments of others in an attempt to clarify contributions and relate them to the general trend of the discussion.	4. Restate comments from members of the group but show no evidence of an ability to tie member contributions together in meaningful relationships.
5. Maintain a high level of work orientation. Initiate work and join others in the task.	5. Tend to avoid the work task of the group and engage in frequent "flight" behaviors.
6. Show a greater facility for maintaining an objective point of view in situations that are emotionally laden.	6. Are more inclined to be influenced by emotionality in such a way that they cannot view material objectively.
7. Are able to make decisions about alternatives that contain emotional elements.	7. Become indecisive in certain instances where the alternatives contain emotional elements.

Of particular interest in the McPherson study is his attempt to develop methods of identifying the high distorters. His primary attempt to do this involved the development of a reading test in which the subject was exposed to paragraphs of material and was later

presented with a series of questions about it. The information provided by the subject in answering these questions was used as a basis for measuring the amount of distortion. Although these results are valuable, it should be pointed out that one may suppose that a person will distort more in making some observations than others. A person will probably distort most in situations in which, for some reason unknown to him, he is motivated to distort.

Unobservables

It has not been found feasible to develop a science of behavior that includes only stimulus variables and response variables pertaining to directly observable events, although traditional education with its emphasis on drill was based to a great extent on the assumption that the amount a child learned was dependent entirely on what was done to him or on what he was forced to do. There was a time when psychologists attempted to develop a science of behavior that required only observables such as stimulus variables and response variables: that is to say, a system in which the laws sought were of the general type $R = f(s)$, where R the response is said to be a function of the stimulus. The character of an individual's response is quite obviously not just a function of the stimuli to which he is exposed, because different individuals show different responses to the same stimuli. This fact can be accounted for only by postulating that individuals differ; or, to use other words, by postulating that different conditions exist and intervene between the stimulus and the response. This is a reiteration of what was said earlier. The concepts are reintroduced here in order to show their relevance to problems of observation.

Intervening variables have sometimes been called hidden variables because they cannot be observed directly, but this fact is easily forgotten in an observation situation. The tendency of observers is to assume that what can be observed—namely, the stimulating conditions and the responses to them—is a sufficient basis for explaining behavior. This is clearly not so except in certain unusual situations that have little relevance for education. Observers commonly make inferences about the operation of intervening variables from the responses they observe. These inferences are not justified. For example, if an observer notes that one child is attending closely to the teacher while another is not, he is likely to make the inference that the one is

highly motivated while the other is not. This statement either merely reiterates that the one attended to the teacher while the other did not, or it invokes the operation of a new variable referred to as motivation. There is no real basis for inferring the operation of such a variable, since the difference in behavior might be due to the fact that the one child was deaf while the other could hear, or that the nonattentive one did not understand English or was dull or was sick. Innumerable other intervening variables can be introduced, all equally tentative and questionable. Without further data, there is no basis for choosing one of these variables rather than another as the correct one. Those variables must be measured independently of the situation in which they operate if they are to be used for explanatory purposes.

Some Problems of Using Untrained Observers

There is much talk in the literature about the need for using trained observers in research, but just how an observer is trained or what this involves is usually left to the imagination of the reader. The problem of training observers can perhaps be introduced by discussing a familiar situation—that provided by a baseball game. The radio commentator observing such a game describes the nature of each pitch, whether it is fast or slow, curved or straight, inside, outside, or down the middle, and so forth. In contrast, the occasional onlooker, such as the present writer, finds it quite impossible to make these discriminations, for all balls look much the same to him and differ only in what happens to them in the subsequent play. The commentator also notices and remarks on many other events that pass unnoticed by the amateur. He notices movement in the outfield as different batters come up to the plate, and other responses of the players to the changing situation. The commentator has, in fact, learned to do two things that the inexperienced observer has not learned to do. First, he has learned to make discriminations that the inexperienced observer has not learned to make. Second, he has learned to respond to more cues, as when he responds to movements in the outfield that few others even notice.

When an observer is said to be trained, it is meant that he is able to make the discriminations required and has learned to respond to all of the elements and relations in the situation in question. When some effort has been made to identify the necessary discriminations

and aspects of the situation, there is no way of differentiating trained from untrained observers.

A professional psychologist is sometimes referred to as a trained observer, but this does not mean that he is trained in all situations. One who has spent his life working with rats on learning problems may be called a trained observer when he is studying the behavior of those rodents, but he may not necessarily be considered trained for conducting an experiment with children. Observation of a particular situation may require special experience that is not provided in ordinary professional training.

In conducting studies that require extensive observations, it is usually necessary for economic reasons to employ observers who have had little training as psychologists. Often graduate students in their early stages of training are selected for this purpose. Under such conditions, it is desirable to follow a few simple rules in the establishment of observational procedures.

First, the observers should participate in the development of the system to be used for recording observations. They should be in on the procedure from the beginning of its development. If this is not possible, then they should be in on the tryout of the tentative schedule. During this process they can be of help in determining what they can learn to observe and what they cannot. They can also acquire facility in the mechanics of recording their observations, and at the same time help in the development of a convenient recording schedule. Through the use of such tentative schedules, observers can compare their records after a period of observation, and, when there are differences attributable to misunderstandings of particular words, come to some agreement on the definition of the terms to be used.

Second, do not expect the untrained observer to record frequently occurring forms of behavior as well as other aspects. He can be kept so busy recording one or two frequently occurring items that he has no time to note anything else. Also, the untrained observer is able to cope with only a limited repertoire of behavior. With practice, this range can be increased.

Third, the schedule not only should specify the category of behavior to be observed but should give the observer training by pointing out some actual examples of this behavior as it occurs in an observational

situation. The categories of behavior must be as specific as possible, and one might possibly suggest the rule that the more naïve the observer, the more specific they should be.

Fourth, categories that involve a considerable amount of interpretation should be avoided. Untrained observers may show little agreement with one another concerning what they consider to be aggressive behavior. Trained observers may ultimately agree on this classification as they learn a common system of interpretation.

Fifth, the observers should be informed of the purpose of the experiment except insofar as this may prejudice the outcome of the study. If the groups or persons to be observed have been exposed to two different treatments and if differences between treatments are studied, it is not desirable to inform the observers of this, lest even a slight prejudice may result in a tendency for the observations to come out in one way rather than another.

Finally, one suspects that if there are great cultural differences in the backgrounds of the observer and the observed, the significance of much that happens may pass unnoticed. This cultural factor is one that scientists are only just beginning to understand.

PRODUCT ANALYSIS AND CONTENT ANALYSIS

It is often convenient to measure the characteristics of a response in terms of a consequence of that response, such as a product. When the product is an object rather than a verbal reaction, its characteristics may be measured; or, if it does not lend itself to easy measurement, a product-rating scale may be developed. The latter type of assessment procedure has been so thoroughly described in textbooks on measurement that it need not be discussed further here. On the other hand, the commonest of human products studied in behavioral research is a verbal product, and the complexity, subtlety, and richness of this material makes the derivation of quantitative assessments a particularly difficult problem. This matter has already been touched upon in the discussion of Withall's technique.

Verbal material represents not only the product of a response but sometimes also a stimulus. Just as we may "observe" the verbal products of a pupil, so too may we observe the verbal materials presented in textbooks, newspapers, and other sources to which he is exposed.

The techniques that may be used for the analysis of the verbal products may be used equally well for the analysis of curricular materials.

The analysis of the properties of verbal material is commonly referred to as *content analysis*, although the same term is applied to the analysis of all forms of symbolic communication regardless of whether it is or is not printed. Content analysis, when it is directed toward textbooks or materials to which the pupil is exposed, may be considered an attempt to identify stimulus variables. When it is directed toward the analysis of the pupil's written products or what he says, then it must be considered to be an attempt to identify response variables.

The history of content analysis, particularly in communication research, has been summarized by Bernard Berelson (1952). Some of the earliest studies pertained to the development of scholarship. Analysis of scientific writings produced at different times over long periods has shown the changing interests of scholars in particular fields and the growth and decline of particular emphases. Other early studies in this field attempted to show the extent to which particular newspapers gave coverage of the news. This was accomplished by defining what was meant by full coverage on particular days by listing the events to be covered. The newspapers were then analyzed to determine what percentage of these topics were covered. Similar studies were also undertaken to compare the reporting of news in different countries, and the analysis of radio broadcasts became an area of military intelligence research. From this there developed a whole area of systematic intelligence research based on analyses of enemy publications, statements by prisoners, and the like.

In education, content analysis of textbooks has long been used as a basis for constructing subject-matter tests, mainly of an informational type. Analyses to determine the more subtle aspects of textbooks have rarely been undertaken, and attempts to identify their subversive aspects have not involved systematic content analysis but usually have been based on a process akin to snap judgment.

Content analysis has had a long history as a means of scoring the products of testing situations. Its uses, and particularly its failures, in the scoring of the traditional type of essay examinations are so well known that they need not be discussed at greater length here. The

reader is referred to other books on the subject if he wishes to pursue further the problems of constructing and scoring essay examinations.

A special case of content analysis, which needs studying but has not been studied, is that involved in the analysis of tests and particularly objective tests. The problem is a common one. Supposing that a teacher or research worker wishes to use a published test for some specific purpose, he may be interested in making an analysis of the test to determine its relevance. In this connection, one group of educators (Bloom, 1956) has suggested that all test items of intellectual achievement be classified into a list of standard categories, which might then become generally used. Examples of the categories they employ are the following:

1.11 Knowledge of terminology
1.12 Knowledge of specific facts
1.21 Knowledge of conventions
1.22 Knowledge of trends and sequences
1.25 Knowledge of methodology
2.10 Translation from one level of abstraction to another
4.10 Analysis of elements
4.20 Analysis of relationships

The above list gives just scattered examples from this attempt to categorize test items in terms of the achievement they are designed to measure. In the source from which these are abstracted, every effort has been made to define each category both in terms of general description and in terms of actual test items. However, despite the care with which this has been done, it is doubtful whether any two educators would show substantial agreement on the classification of items. The reliability of the process of classification needs to be determined experimentally.

This matter is closely tied up with the whole problem of determining the content validity of an achievement test. The term "content validity" is here borrowed from current usage. It is really inappropriate, since it refers, or should refer, to the objective properties of a test item. The so-called content-validity problem is generally stated as that of determining the extent to which a group of test items can be considered to be a representative sample of a universe of items. This can be determined only if the dimensions of the universe are

known. This problem of sampling is important in all fields of content analysis.

An aspect of content analysis that has had great significance for educational research is that related to the appraisal of the comprehension level of printed material. A whole series of readability formulas have been developed, but there is still much controversy concerning their appropriate uses and limitations. The earliest attempts to measure this characteristic of printed material used the simple expedient of determining the percentage of difficult or easy words in terms of a list giving the frequency with which words appear in certain types of published materials. Those who undertook these early measurements were aware that this simple procedure was inadequate for handling the complexities that even the most straightforward prose presents. It soon became evident that these simple methods had very little use, but not until the mid-1930's had sufficient research been undertaken to permit the construction of more useful methods. Gray and Leary (1935), who pioneered in this field, published a complex formula for measuring readability, based on five characteristics that had been selected from a list of eighty-two assumed elements. Nearly a decade later, Lorge (1944) developed a similar formula based on only three characteristics—the number of hard words (as determined by a standard word list), the number of prepositional phrases, and sentence length. Other formulas appeared at about the same time as the Lorge formula; they included the Flesch formula (1951) and the Dale and Chall formula (1948).

There seems to be considerable agreement that the characteristics measured by these formulas are not entirely satisfactory and that the formulas cannot be applied to all kinds of materials. A formula that is quite satisfactory for measuring the readability of grade school material may provide ridiculous results when it is used for assessing the readability of technical writing. A central difficulty is that the formulas do not take into account some of the more subtle aspects of style, which may have enormous influence on the difficulty level of reading material. An author may write in simple words, but his material may be difficult to read because he makes use of unusual analogy and innuendo in a way that conveys a richness of meaning through its overtones. No formula at present available takes into account this aspect of reading difficulty.

Reading difficulty is also a complex function of a person's previous experience. An elementary textbook in physics may be very difficult for a student before he has taken a course in physics but easy once he has mastered the vocabulary and concepts of the field. Relatively slight differences in life histories may produce marked changes in readability of material. This fact illustrates the weakness of readability as a concept. When readability is being measured, the scientist is not actually measuring an objective property of certain stimulus material; rather is he making a prediction of how individuals on the average will respond to this material. He is measuring, in effect, a response-inferred stimulus property, which is a rather weak variable to fit into a matrix of scientific ideas.

Measures of readability have many applications, both in school problems and in the broader field of public education. They provide an objective means of determining the suitability of materials for various age groups and for various levels of pupil ability within these groups. They provide a means of adjusting existing materials to a more suitable level of readability, and many items of classical literature that are quite inappropriate to the reading level of pupils have been adjusted to make them suitable through the medium of readability formulas. In addition, such formulas have been used to measure the readability of announcements and other materials designed for public education programs such as are sponsored by various health and safety organizations.

Content analysis of the type thus far considered presents special problems of sampling, because it is not usually feasible to make an analysis of all of the material available. In the preparation of word lists indicating the frequency with which various words appear, the problem is acute. The purpose of such counts is to measure the "difficulty" of words, but the term "difficulty" has meaning only when it refers to a particular individual or group. A word is "difficult" if it is difficult to understand or if its meaning is understood by only a fraction of the members of a group. Thus if the difficulty of words for a particular group is to be estimated from their frequency in reading material to which the group is exposed, it is necessary to know what this group reads and to sample it. It may happen that the members of the group differ greatly in what they read, in which case it is necessary to assume that at least some of these various

materials are equivalent. Of course care will be taken to insure that all the materials on which word counts are made are not by the same author, because individual authors often have their own favorite vocabularies. The difficulty of words assessed from a representative sample of specified materials may not provide a useful estimate of difficulty of words for other groups that have different reading habits. The common practice in this area is to sample so-called "popular" reading materials, such as the *Saturday Evening Post,* in the hope that the results can be applied to a wide range of groups, but this method of identifying the sample to be analyzed has always been criticized.

Content analysis at more complex levels than that of word difficulty ceases to be an entirely objective matter. In the analysis of printed materials, the analyst is faced with a series of black marks on a piece of paper. The marks he must interpret, and the interpretation process involves the same subjectivity as does the interpretation of pupil behavior in the classroom. There should be agreement among judges concerning the interpretation of materials before anything more than the simplest content analysis can be made. This is fairly easy when the analysis is at the level of counting words in different categories, since there is good agreement as to which words are to be classified as nouns, which as adjectives, which as verbs, etc. Greater difficulty is experienced in separating factual information from inference. Perhaps little reliability might be found if an attempt were made to single out from a total speech those remarks that reflected hostile gestures. We can think of a continuum that varies from one end of the scale, where there would be no agreement among analysts, to the other end of the scale, where there would be complete agreement. Under the latter conditions there is objectivity of measurement, a term meaning that there is social agreement. It is not unusual to find that those aspects of content analysis on which there is complete social agreement are the most trivial of those it is desired to measure. It is thus necessary to move further down the scale and to sacrifice some objectivity in favor of relevance.

As thus far considered, content analysis has not dealt with the classification of the *ideas* portrayed by written or printed materials. The latter type of content analysis has had a long history, but only recently does it seem to have taken a profitable turn. It finds its roots

in some of the early free-association experiments, in which subjects responded to statements with the first word that came to mind. Attempts were made to classify and analyze the responses to obtain indications of interests, attitudes, the nature of repressed ideas, and a host of other conditions. These early attempts were not particularly successful, and indeed they were sufficiently negative to discourage a generation of psychologists from further pursuit of the method.

Reference is made here to the type of free-response test that has become known as a personality test, in which a stimulus is provided and the response is scored in some way that is believed to be relevant to the broad description of behavior. Such attempts to measure personality have a long history. Tests in which responses to ink blots are scored have a history that goes back before the time of Rorschach. The tendency in the early days was to score in terms of content categories, such as animals, plants, tools, structures, and buildings, etc. The failure of these categories to produce useful results tended to discourage investigators from using this approach. Word association tests also produced disappointing results. In their pioneer attempt to score objectively free associations to words, Kent and Rosanoff produced a method of classifying single-word responses, but the instrument with its complicated scoring system failed to achieve the practical results expected of it.

The early period of discouragement with content analysis of the responses to personality tests was replaced by new hope through the work of Rorschach and his followers, who realized what others had not realized—namely, that the analysis of responses could be made along numerous dimensions. They saw that it was not necessary to stay with subject-matter categories, but that other aspects of the response could be scored. Interpretations of ink blots could be scored in terms of whether they referred to the blot as a whole, a major part of the blot, or a minor detail. They could be scored in terms of whether they were responses to shading, to color, or to the outline of the blot. They could be scored in terms of whether they referred to stationary objects or to moving objects. Rorschach himself seems not only to have had ingenuity in the setting up of such categories but also to have realized the importance of basing them on a rationale. Whatever success the Rorschach test may have had seems to have been largely a result of his talent in selecting scoring categories with

a rational relationship to the categories of behavior they were designed to predict. The difficulty, of course, has always been in the establishment of the relationship between response category and the aspect of behavior that it is desired to predict. However, although the actual evidence available for such a relationship is slender, clinicians have felt that the Rorschach categories offer promise, and largely on the basis of their opinions, categories of this type have become widely accepted as perhaps the most promising method of content analysis yet devised.

The types of categories used in scoring the Rorschach became a prototype for scoring projective tests for a period of nearly twenty-five years. Recently, however, there has been a revived interest in the scoring of responses in terms of categories of objects or events. This newer emphasis may be traced mainly to the developments made in the late 1930's by H.A. Murray of Harvard and his associates, which sparked a whole series of important developments that are now beginning to yield results of great practical importance. Murray postulated that human motivation could be measured in terms of a set of needs, such as need for achievement, need for sex, need for autonomy, and need for affiliation with a group. He also postulated at the time, although the idea was extended and developed at a later date by his followers, that such needs could be measured through projective tests. The immediate outcome was a test known as the Thematic Apperception Test, which consisted of a series of pictures deliberately drawn so that the situations portrayed were somewhat ill defined and the human figures were vague and ephemeral creatures. The pictures were shown to subjects, who were asked to interpret them, and since they presented indefinite stimuli, the interpretation had to be a structure imposed on the picture by the individual describing it. The structure could be considered to some extent a product of the personality of the observer, reflecting his needs, motives, salient characteristics, and so forth. The Thematic Apperception Test, which became known as the T.A.T., never yielded a satisfactory system of classifying the responses in terms of their content and thus never yielded a satisfactory scoring system, at least not until parts of it were adapted by McClelland and his associates and a new notion of idea analysis was introduced.

The derivation of objective measures from instruments such as the

Thematic Apperception Test presented problems that it has taken the best part of two decades to solve. The protocols derived from the test presented a range and richness of material that for a time seemed to defy objective categorization. Most of the early attempts to score such material resulted in elaborate systems that proved to be much too cumbersome to be of any real practical value, either in the development of research or in the assessment of personality in practical situations. The Tomkins (1947) method of scoring may be reviewed by the student as an illustration of one technique that is as ingenious as it is complicated.

The first major step in the overcoming of the difficulties involved in the scoring of such material seems to have come with an important observation by McClelland and Atkinson (1953) that subjects deprived of food for varying periods of one, four, and sixteen hours showed increasing numbers of references to food, food-getting, and hunger in imaginative stories that they were asked to produce. The observation is to some extent a repetition of that made nearly twenty-five hundred years earlier by Buddha, who observed that fasting failed to clear the mind of worldly things so that it could dwell on higher values. To the contrary, he found that the well-fed state is most conducive to the study of philosophical and ethical problems. The reader will note that this is the basis for presenting Buddha in statuary as a well-fed figure, if not as one who has eaten well beyond his physical needs.

The rediscovery that deprivation of whatever satisfies a need results in an increase in the amount of imagery related to that need opened a new era in projective measurement. McClelland also realized that it was desirable not to use just any picture in the arousing of imagery related to a need, but what was required was a stimulus object that would facilitate the arousing of such imagery. Thus the pictures used in the original Thematic Apperception Test did not appear to be particularly appropriate for measuring need for achievement, in which McClelland was interested, and he found it necessary to construct pictures more appropriate for this purpose. One of these shows a boy seated beside a table (or desk) on which there are a number of books. The boy is leaning his forehead on his hand and looking upward as if in reverie. (The writer hopes he is not projecting too much of himself into the interpretation.)

The person who takes the McClelland test is asked to answer a series of four questions about the picture, which are:

1. What is happening? Who are the persons?
2. What has led up to this situation—that is, what has happened in the past?
3. What is being thought? What is wanted? By whom?
4. What will happen? What will be done?

The answers to these questions may vary from a few words to what amounts to a long essay. It is not the volume of material produced that is of interest to the psychologist, and scoring systems should not be related to mere quantity.

The trend in the scoring of such material is to score each statement in terms of a number of categories. One category commonly used is the *desire for a goal*. If the subject writes, "The boy in the picture is determined to get a high grade," the achievement imagery represented by this response would be scored in this category. If the same subject were to have written, "The boy is thinking through his work carefully to get the highest grade he can," it would have been scored not only in the *desire for goal* category but also in the *goal-directed activity* category. Other categories that have been used are those referring to *expectation of success, expectation of failure, desire to avoid failure*, etc. A purpose of scoring each recorded item in several categories is to derive as much information as possible from the material provided and thereby to increase the reliability of particular scores.

The type of approach that McClelland has opened provided a means of scoring a great range of material. For measuring achievement and affiliation motivation, French (April, 1956) developed an instrument that used a closely related scoring system but applied it to somewhat different material. Her test consisted of a series of brief descriptions of behavior, such as, "Ray works much harder than most people," and, "George will usually volunteer for a difficult task." The subject is asked to indicate why the person described behaves as he does, and the material is presented as a test of insight into the behavior of others. Since one cannot know the underlying cause of the behavior described in the test, the interpretation given must be largely a projection of the individual's own motives.

These scoring procedures have been successful in producing vari-

ables of consequence; that is to say, the variables lead to the making of useful predictions and produce at least low-level laws that are consistent with other knowledge and with expectation. They represent definite advances over earlier procedures. The scoring of the Thematic Apperception Test type of instrument has taken nearly twenty years to evolve. This has been a slow process, but scientific advances are typically slow. Perhaps the moral to be drawn is that the essential element for the development of any sound measuring instrument is a sound theory concerning how and why a particular procedure will evolve a useful device. If the theory is sound, it is probable that a useful measuring instrument will result.

The chief disadvantage of this type of measuring instrument is that the scoring procedures are extremely laborious. For this reason, the procedures must be considered as tentative, and they will probably be replaced ultimately by a measuring procedure much less laborious. This usually involves the substitution of checking procedures for free-response procedures. So far, no successful substitution of this kind has been made in the case of projective techniques.

The secret of the success of McClelland's type of content analysis seems to lie in the fact that the stimulus is such that it restricts the range of the response, and it is this restriction that makes the material so amenable to analysis. A major reason for the success in the making of predictions of the scores thus derived is probably the fact that the total procedure is based on what is, within limits, a sound theory of motivation.

In a sense, the content analysis of McClelland and his associates represents a revival of an earlier psychological movement that had attempted to build a science of the content of the field of consciousness. However, it is a revival that seems to be accompanied by much greater success than its earlier versions.

Projective techniques and the content analysis needed to derive scores from the material may be specially adapted to the study of educational problems. For example, Figure IV represents an interesting approach to the study of teachers' concepts of their role in a classroom. Teachers were asked to draw a picture of a teacher with a class. It is believed that the products indicate the teacher's own concept of how he should behave in a classroom, and there is some evidence to substantiate this point of view.

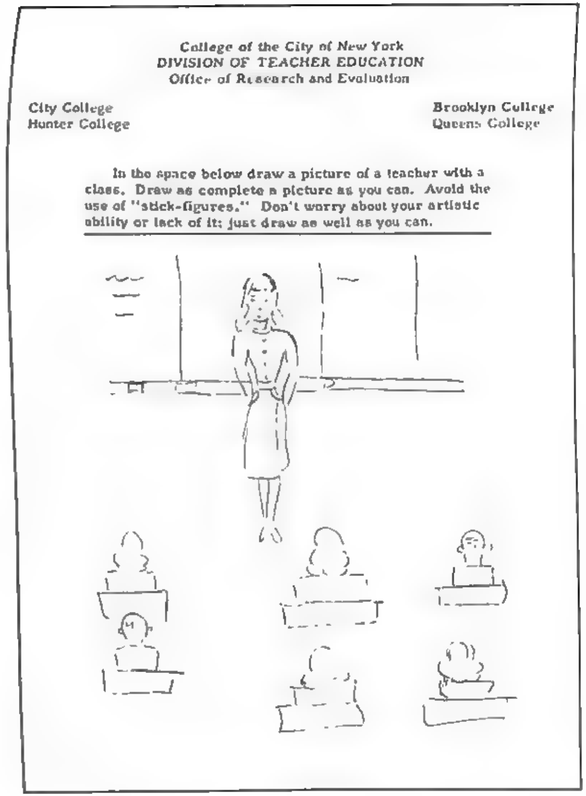

Figure IV. A projective technique for studying concepts of teaching. This illustrates the use of a projective technique in which students of education and teachers were asked to draw a picture of a teacher with a class. This illustration presents two strikingly different concepts of how a teacher should behave in a classroom, and also how a classroom should be organized. Illustration by courtesy of Dr. William Rabinowitz, and collected as part of a research study at the Office of Research and Evaluation of the Division of Teacher Education, Municipal Colleges of New York City.

ADDITIONAL OBSERVATIONAL TECHNIQUES

The Critical Incidents Technique

During the present decade, interest has developed in what is called the critical incidents technique, an invention of John C. Flanagan, who has applied it to a great range of situations. It is, in essence, a method of observation, but it also involves the judgment of the observer concerning what should be observed and recorded. The technique is a method of defining the group's concept of what makes a particular member of an occupation more effective than other members. It is a way of defining, for example, what superintendents mean when they say that teacher X is more effective than teacher Y. It avoids the difficulties that are produced when a superintendent is asked the question, "What do you consider to be the characteristics of a good teacher?" The answer to such a question always involves vague generalities, such as, "The good teacher is kind," or "The good teacher provides effective incentives." Such descriptions are not a sufficient basis for developing instruments that can be used to measure the extent to which teachers do actually conform to this ideal. The critical incidents technique may be used to describe more adequately what a superintendent has in mind when he states that teacher X is better than teacher Y, but this does not in any way mitigate the fact that the technique is still operating in the domain of judgments. In the ultimate analysis, the pronouncements of any person concerning what makes a good teacher must be based on a reasoned judgment that *this* type of pupil learning is better than *that*.

The procedure involved in the critical incidents technique is well illustrated by a study by Jensen (1951) of the critical requirements for teachers. In the Jensen study, a *critical incident* is an observed teacher behavior or aspect of teacher behavior that is *judged* to make the difference between success and failure in teaching. The term *judged* is italicized here to indicate that what is being accomplished is to summarize such judgments in terms of behavior incidents. There is clearly little point in collecting descriptions of the commonplace behavior that is about as typical of those teachers judged to be good as it is of those judged to be poor. The reader must also note that the method is not a scientific device for deciding what constitutes

good teaching. It is only a method of describing what some person or group considers to be crucial matters in judging the merits of teaching.

In the Jensen study, the participants were asked first to recall an elementary teacher judged to be ineffective, and then to relate the incident that made the participant decide that the teacher was incompetent. Similar questions were asked concerning the participant's experiences with effective teachers. Jensen also tried a number of variations of this technique, such as asking the participants to go back to their own childhood and recall incidents of effective and ineffective behavior on the part of teachers.

The material derived from such a technique is voluminous and needs to be reduced to manageable proportions. This may be done in various ways. One of these is to attempt to abstract from each incident the salient feature that caused it to reflect effectiveness or ineffectiveness. Thus one category might be "overcriticalness" on the part of the teacher, another "fairness and impartiality." One can then compare frequencies between classes, or compare groups of teachers with respect to the frequency of occurrence of any particular class of behavior. In addition, one may use the material for constructing scales for rating teachers. The technique could be applied to the definition of a "good" student versus a "poor" student or a "good" parent versus a "poor" parent, and to almost any situation in which one group is to be discriminated from another. Nevertheless, this statement concerning the wide applicability of the technique should not be taken to mean that the results are necessarily useful.

The critical incidents technique has enjoyed a period of relatively uncritical popularity, which is quite typical of new techniques in an area where research workers have often a feeling of helplessness because of the complexity of the problems faced. The technique also has an apparent attractiveness and relevance to the solution of many important problems, though it is the opinion of this writer that the attractiveness is superficial. It is, therefore, necessary to examine the technique rather more critically.

First, the technique often provides samples of rarely occurring behavior that are not likely to be observed again. The very rarity of the observations minimizes their usefulness.

Second, the infrequency of the events listed in any study of critical

incidents makes them extremely difficult to classify. The event that is remembered and later recorded as a critical incident is often remembered because it is unusual. Thus the technique reflects the phenomena of selective recall and often produces what appears to be a list of unique, and therefore unclassifiable, events.

In addition, the technique may reveal substantial lack of agreement among participants. If this occurs, the product is likely to be a conglomeration of events that cannot be shaped through categorization into any recognizable form. Even if there is some agreement, it may not be possible to arrange the behaviors along any kind of continuum, which would be a most desirable feature of the data if they were to be used as the basis for a check list.

Finally, the technique should be recognized by the graduate student of education as an extremely laborious one. It involves the manipulation of large masses of data through techniques that are extremely time-consuming. It should not be considered as one suitable for a doctoral dissertation or a master's thesis. Even if the difficulties involved in the handling of the volume of data are overcome, there would still be considerable doubt that the lengthy and tedious work involved would yield results of consequence. Although the technique has been applied to several types of educational problems, results have not been of any particular note.

Self-Observation and Self-Report

The oldest technique of the psychologist is that of self-observation, and it still occupies a place in modern psychology. Self-observation has undergone a long history of development, and current techniques that might be considered to fall in this category are far removed from those that would have been used a century ago. Much of what is strictly self-observation is also not commonly referred to by that term and is consequently not recognized as such.

Self-observation techniques began in history with crude introspection of the armchair type, whereby through an inspection of one's inner self an effort was made to discover the laws of events in consciousness. This type of self-observation acquired some refinement in the first psychological laboratory founded by Wundt and reached its highest stage of development in the work of Titchener and his pupils. Despite all the work devoted to the development of the intro-

spectionist approach, it is viewed today as having been a profitless venture. Even though the techniques evolved may have done much to eliminate errors in the observation of mental phenomena, the knowledge acquired by these means was meager. Today it would be hard to find a laboratory in which these techniques would be considered for use.

One plausible reason for the failure of classical introspection techniques to yield useful results was that they did not involve measurement. More recently attempts have been made to quantify certain aspects of personal experience, and the results of such efforts seem promising. Certainly it can be said that advances in the behavioral sciences seem to have accompanied the introduction of quantitative methods. With this in mind, a brief review will be made here of quantitative self-observation techniques.

Self-ratings. Graphic rating procedures have been applied extensively to the situation in which the person rates himself. The early studies of this type were based on the concept that the individual is the person who knows himself best, and that self-ratings could be used as a basis for study of the structure of personality. This concept had some value, naïve though it was. R.B. Cattell, who has worked extensively in the area of personality ratings, is of the opinion that self-rating studies help to confirm findings based upon other sources of information. In the last two decades, self-ratings have come to be viewed in a rather different light as indicators, not of a personality as it actually is, but of a person's self-concept. The distinction between a person's self as it appears to others and as it appears to himself has been found to be a useful one. Psychologists of the Rogerian school have emphasized that a person's concept of himself may have a powerful influence on behavior. Reference is made here to the school of psychology that emphasizes the importance of events in the phenomenal field. According to this viewpoint, the development of a science of psychology involves the discovery of the laws of the sequence and interrelationship of events as they occur in the person's field of consciousness. This field is described as the *phenomenal field*, a term that has been handed down from German philosophy, where it was used to denote the field of consciousness within which all observable phenomena occurred. Phenomena were events and things in the universe as they were observed,

and they were contrasted with noumena, which were things and events as they actually existed independent of an observer. According to the phenomenologist, all the determinants of behavior exist in the phenomenal field. Therefore, in order to understand behavior, it is only necessary to study the phenomenal field. Since the psychologist does not have direct access to the phenomenal field of another, he must rely on the statements made by that person about his own phenomenal field. The statements that a person makes about himself are also used by the psychologist who has a behavioristic outlook. He regards such statements as objective facts and ignores any reference they may have to an inner life.

Q-methodology. An extension of self-rating techniques with special methods for evaluating the data is represented by the type of methodology that Stephenson (1953) developed and named Q-methodology. His ideas offer some promise of permitting the study of the structure of self-concepts or of personalities as they are perceived by others. During a period when he was visiting professor at the University of Chicago, he was brought into close contact with the conceptual system of Carl Rogers and became intrigued with the problem of assessing how a person's self-concept changes during psychotherapy.

The methodology that Stephenson had developed several years previously seemed to be particularly well adapted to this purpose, since it permitted over-all comparisons between persons or between assessments of the same person made at intervals of time. Several studies were undertaken that demonstrated the feasibility of showing changes in the self-concept as psychotherapy progresses.

Personality inventories as records of self-observation. Another method by which personal experience may be quantified is through personality inventories, which are mainly devices for recording personal experiences. Some of these devices limit such personal experiences to responses of liking and disliking and are referred to as *interest inventories*, while others are restricted to different categories of personal experience. Still other inventories deal with the course of action that a person would most likely pursue when faced with certain kinds of situations. Another entirely different approach to the quantification of personal experience is through the *biographical inventory*. Devices such as the latter attempt to quantify certain aspects of a person's background as it is revealed through his ex-

perience. One can include in biographical inventories items for which the responses can be objectively verified, but such items are likely to be trivial and of less consequence than those that are not verifiable.

It is difficult to provide an appraisal of such an extended field of measurement in the short space that can be devoted to it here. The measures derived by such self-observation techniques have not shown the promise that was originally anticipated for them. However, they have not shown themselves to be entirely useless, for situations have been found in which they have provided predictions of limited accuracy, but there has often been difficulty in reproducing these predictions in related situations. Interest inventories have perhaps been the most widely used of these instruments. In comparison with the recent success achieved with some types of projective techniques, the success of the inventories considered here has been meager but not entirely negative.

Self-observations have come under criticism from two quite distinct standpoints. First, it must be pointed out that clinical psychologists have emphasized since the days of Freud that a person is an extremely biased observer of his own behavior. Much behavior of great significance escapes his observation, and he is a poor judge of what are the significant and what are the insignificant aspects of his behavior. In addition, many of the important determinants of behavior are not considered by the intellectual descendants of Freud to be within the realm that is open to personal and direct observation. Motives, for example, are said to be in the category of unobservables in the person who has not undergone psychoanalysis, although the latter therapeutic technique is said to provide some personal access to this aspect of personality. Second, the person who is observing himself is a biased observer. He wants to present himself to others in the best possible light, and his answers to questions about himself are colored quite inevitably by this tendency. The person observing himself is not untruthful in his report, but is he selective in what he reports and in how he reports it. He may at times report in a way that is truthful yet biased.

The Utilization of Biographical Data

Biographical data have always been of immense interest to the educator. Every teacher recognizes that the problems and difficulties

of children are, at least in part, the product of previous conditions. Most of us have the implicit belief that if we could know the complete life history of an individual, then we would understand his present behavior. The teacher, the counselor, the social worker, all seek information about the past history of the individual in order to understand his present actions. The research worker interested in such problems may attempt to obtain data about the past in order to determine how present behavior came into being. Since biographical information is one of the more important classes of observations that are made in educational research, some consideration needs to be given here to the data-collection problems that it involves.

In the traditional type of biographical study, biographies are examined on an intuitive basis much as the clinician examines a patient. An example of this type of approach is manifested by Anne Roe in her studies of creative talent. The purpose of the examination of the biographies in this case is to determine whether the group of creative individuals show any common characteristics running through their lives. In the case of Anne Roe's studies the attempt seems to have had some success, and the results of the biographical studies have been confirmed by other sources of evidence. Nevertheless, the success of this method in the case of some studies does not mean that it is always successful. The truth seems to be that the method has many dangers, and unsuccessful applications tend to be overlooked.

A major danger is illustrated by the early biographical studies of neurotic patients. In these it was shown again and again that such individuals often had been exposed to traumatic experiences. The conclusions were erroneously drawn that traumatic incidents in childhood produce neurotic behavior in adult life. This conclusion is not justified, for when the background of well-adjusted individuals is also examined, it is found that this group too shows a similar incidence of traumatic events. A related error was made at one time as a result of investigations of the family background of psychotic patients. It was found that such patients had a large number of relatives who were "queer" in some way. It is not reasonable to conclude on this basis that psychoses are inherited, since further investigation shows that so-called normal individuals also have numerous relatives who are commonly described as "queer." The reader will recognize that the way to prevent such erroneous conclusions is to introduce a group

of "normals," whose background is also examined. The introduction of a control group is really necessary in order that any conclusions at all may be reached.

The biographical information presented by autobiographies or derived from interviews is difficult to treat in any scientific study because of its diffuse nature and because of the multiplicity and variety of the events that it may cover. These characteristics force on the investigator the intuitive approach that must be taken in examining such material. The intuitive approach involves the interpretation of the material, but an interpretative process invariably introduces error. In order to avoid such errors, inventories have been developed for recording biographical information.

In the typical biographical inventory, a standard series of questions is asked about a person's background. The questions are answered by choosing one of a number of alternatives. Typical questions are the following:

In what type of community did you spend most of your time before entering school?

1. In the country
2. In a town with less than 5000 inhabitants
3. In a town with 5000 to 10,000 inhabitants
4. In a town with 10,000 to 50,000 inhabitants
5. In a town with more than 50,000 inhabitants

Which group of school subjects did you prefer when you were in high school?

1. English, speech
2. Social studies, history, geography
3. Science, mathematics
4. Music, art
5. Athletics

Biographical information collected in the form illustrated above has had a long history of practical use and also some history of having played a useful part in research. Many have regarded it with skepticism, for reasons that still have to be considered, but the fact that it has had a long history of practical utility in the selection of various classes of employees has forced researchers to give it serious

consideration. It is of interest to note that the first really successful use of the biographical inventory was in connection with the selection of salesmen, and particularly life insurance salesmen. Such devices remain, even today, the main instruments that are used for this purpose. Of course, such devices were not developed on the basis of any particularly sophisticated psychological theory of selling. The point stressed here is that this work of practical importance demonstrated that biographical information collected in this form could be used for making predictions, and probably with more success than biographical information collected in narrative form. Of particular significance is the fact that biographical items related to factual material had considerable predictive significance, while those related to opinions and attitudes tended to be of doubtful value.

During World War II some success was achieved in the use of biographical information blanks for the prediction of performance in flying training, and there were even indications that they could be used for predicting aerial combat leadership. These later inventories that appeared in the wartime program were much more sophisticated than earlier devices in the theory on which they were based, and this influence has been shown in work that has been undertaken since that time. A major development incorporated into more recent biographical information blanks has been an attempt to group items in such a way that they measure a number of distinct and separate influences in a person's background, or even a series of relatively independent traits that may emerge from such backgrounds. There has also been considerable interest in attempts to predict variables other than occupational success. For example, there have been many studies in which biographical information has been used to predict reaction to stress, and predictions of sufficient accuracy to be used have been achieved.

The clinician has never been particularly in sympathy with this approach to the matter of using biographical information. He has tended to feel that the very uniformity of the material included in a biographical information blank is a disadvantage. He points out that the unique event is often a crucial factor in the life of an individual, and this would be missed by any standard inventory. The clinician has not proved his case in this matter, and the success achieved with biographical inventories may perhaps make him stop and ponder.

Summary

1. The observation of behavior in the classroom and in other complex situations usually requires the use of observation schedules.

2. Observation schedules should include no more items than the observer can remember and easily locate on the list. The schedule should usually refer to items of behavior that occur with fair frequency. It should also be easy to recognize when an item of behavior has, or has not, occurred.

3. Observation schedules should be based on a theory concerning the relevance of the items of behavior observed to the purposes of the study.

4. The student should be on guard against a procedure for selecting items that results in the selection of those that are highly reliable but rather inconsequential. A check-list type of schedule may be quite unsuitable for the purpose of assessing the more subtle aspects of classroom phenomena, and the latter may have to be appraised in terms of the general impressions of the observer. The latter is a practice that research workers have strenuously attempted to avoid.

5. An interesting proposal is that observation be directed toward the verbal communication that occurs in the classroom. Since most of the transactions that occur between pupil and teacher are undertaken in terms of words, it seems reasonable that the analysis of these words will provide significant information about events in the classroom.

6. In most observation schedules an attempt is made to measure certain dimensions of behavior, rather than merely to keep a record of how frequently this or that event happened. Research workers are still exploring the usefulness of various systems of dimensions that may be used for describing events in the classroom, and one cannot state at this time that one system is more useful than another for particular purposes.

7. Attempts have been made to introduce mechanical devices as means of simplifying the observer's task.

8. Research has indicated that some persons are much more prone than others to introduce distortions into their observations. The high distorters have characteristics that differentiate them from low distorters.

9. Any system of observing behavior must take into account the fact that there are unobservable conditions affecting behavior.

10. The problem of training observers was discussed. The training process involves the acquisition of the ability to make certain kinds of discriminations with speed and accuracy. If those who are to observe are able to participate in the development of the research, it is likely that this will help them to learn the discriminations to be made.

11. The analysis of the content of verbal behavior is an important

adjunct to the other types of observational techniques discussed. The content analysis of textbooks and tests is an activity commonly needed in education. Formulas for measuring readability and other devices may supplement the more obvious techniques, but such formulas are not entirely satisfactory.

12. All content analysis involves problems of sampling. It is rarely possible to make an analysis of all the verbal material available, and hence it is necessary to draw samples from it. The samples must be such that their analysis will provide results that apply to the whole of the material.

13. In recent years some success has been achieved in the analysis of some of the more subtle properties of verbal material, such as the type of imagery that it indicates. McClelland's techniques represent major advances in this area.

14. The critical incidents technique represents a special method for selecting observations in certain situations. The technique has certain difficulties associated with it, which should be recognized.

15. Techniques have been developed for systematizing the process of self-observation. These techniques include self-rating procedures, Q-methodology, and personality inventories.

Some Problems for the Student

1. Attempt to build a list of *rewarding* behaviors manifested by teachers toward the children in a classroom situation. This check list is to be used as a means of recording classroom observations of teachers.

2. An interesting exploration in content analysis is to make an analysis of the stories found in children's readers in terms of imagery of the type discussed by McClelland. Try to determine whether the imagery reflects predominantly achievement motivation, status motivation, or other motivations. Many doctoral dissertations could be undertaken in this general area, but the proposal here is for the student to explore the application of McClelland's techniques for the analysis of reading material provided in schools.

Survey Methods | 10

Surveys are conducted to establish the nature of existing conditions. A school survey is commonly conducted in order to determine the services that a school can render a community and perhaps to compare these services with those that are provided by other schools.

A distinction must be made between a survey and a sample survey. In a sample survey, data are collected about only a portion of the events with which the surveyor is concerned. The design of the sample survey is such that the events examined provide data from which inferences can be made about all of the events. Thus an examination of one hundred rural high schools within a particular state may permit the making of inferences about conditions in the other five hundred.

The survey method represents research at a primitive level. It builds a body of fact that is usually of only local significance. The facts thus collected may contribute to the solution of immediate problems, but rarely do they develop a body of knowledge that can be used in solving future problems. Thus the technique tends to be a one-shot method.

The survey approach to educational problems is one of those most commonly used. It is followed in studying local as well as state and national problems. It is the commonest method used by principals to obtain information about some problems. Most surveys result in the analysis of verbal responses, such as those given when an interviewer asks a question and obtains a reply. Similar methods may also be applied to the survey of concrete objects or other types of conditions and events. The Research Department of the National Educational Association has devoted a great deal of its effort to the survey of such matters as school transportation, size of classes, and various aspects of the physical facilities of schools. There is, however, a second type of survey conducted in a school or a school system, which is rather different from the type of survey that has been considered. This is the type of survey conducted by a consultant who has been called in to cast an expert eye over a school system and to advise on the points where the school or system needs to be improved. In this type of survey, the expert collects a great range of facts about a single school or system. He is interested in every phase of it, and no detail escapes his regard. He is concerned with putting together these details in order to provide a total picture as seen by an objective outsider. His approach is qualitative and does not lend itself to statistical treatment, for it is not built around hypotheses that can be tested by typical statistical methods.

The events or conditions that may be enumerated or measured during surveys include a great range of phenomena. The popularity of the public opinion survey as a newspaper feature makes one think of attitudes and opinions as the main source of data in surveys, but this is not necessarily so. Various classes of events that may form the central core of an educational survey must be given brief mention here.

Physical conditions related to learning. Many characteristics of the physical environment may be measured, such as the floor space per pupil in different schools, the number of books per pupil in the library, the intensity of the natural illumination on the pupil's desk, the mean temperature or humidity, etc. Many school surveys devote considerable effort to the measurement and evaluation of such characteristics of a school program, and on the surface they would appear to be on a solid foundation, for it is clear that the *measurement* of

these variables is objective and does not involve the judgment of the investigator. These measures have satisfactory reliability, and thus their weakness is not at first apparent. The inadequacies of the procedure are a result of the fact that its usefulness depends upon the choice of suitable environmental variables—that is to say, variables that are genuinely related to the effectiveness of learning of the pupil. But very little is known concerning the relationship of such variables to learning processes in pupils. If the student looks back over research in the area, he is likely to find not only little positive evidence to help in the selection of variables but much negative evidence indicating the apparent lack of relevance of many variables he might choose. Perhaps such studies have failed to demonstrate the relevance of some of these physical variables because any effect they may have is perhaps long-term and not sufficient in magnitude to manifest consequences over a period of a few weeks or even a semester. Long-term studies of the effect of these physical conditions are rarely feasible.

Behavior of teachers and other behavioral conditions related to learning. Surveys related to the behavior of those who control the educational process are sometimes undertaken, but they involve difficulties of which the prospective researcher must be aware. Many of these surveys pertain to the problem of assessing teaching effectiveness and are based on the assumption that particular characteristics of the behavior of the teacher facilitate learning in the pupil. Rabinowitz and Travers (1953), who studied this problem, pointed out that it is quite obvious that such a simple relationship cannot exist. Even the casual observer in the classroom must have observed that some teachers are aggressive and tend to induce fear, while the same aggressiveness in other teachers, combined with other different qualities, produce enthusiasm and high motivation for work. It also seems probable that teachers are effective in different ways. Some are effective in explaining and demonstrating, while others have special skill in organizing socialized activities. One may presume that such differences in skill may produce differences in the development of the pupil, though just what they do has not yet been determined. Also, it seems probable that a teacher may provide a favorable learning situation for some pupils but not for others. There is a complicated type of interaction between teacher characteristics and pupil char-

acteristics, which has not yet been properly described and which cannot be described for a long time to come. For these reasons, surveys of teacher characteristics, made for the purpose of discovering the extent to which conditions favorable to learning exist, assume knowledge that has not been established. It is therefore highly doubtful whether much can be accomplished at this time by such surveys.

Some investigators who realize the many weaknesses of describing characteristics of teacher behavior, but who have wanted to determine the extent to which conditions favorable to learning exist in the school, have conducted studies on the adequacy of teacher preparation, and have based these studies on the assumption that the better the preparation, the better the teaching. It is one of the very embarrassing facts of teacher education that we do not as yet have any direct evidence to show that this assumption is true.

More justifiable are surveys of opinion of various groups designed to provide a basis for establishing objectives or goals. Students of education know that objectives of broad educational programs are necessarily based on the value judgment of some person or persons. The selection of objectives is generally considered to be best undertaken as a group decision rather than as an individual decision, and hence surveys represent appropriate techniques for this purpose.

The results of learning or the ability to learn on the part of the pupils. The surveys that are most likely to reveal facts of importance for educational administration are those provided by the pupils themselves. Such information has a direct relevance to the control and study of the learning process, which the other classes of facts do not have. Under this category surveys may be made regarding the reading achievement of pupils or their achievement in other so-called basic skills. Sometimes surveys of the information of the pupils may be made, as when a school determines what the pupils know or do not know about their local community, about health practices, about contemporary affairs, or about some other matter judged to be of significance in the educational program. Such surveys need not be confined to matters of student knowledge but may also include events in the attitudinal field.

Much survey research conducted in this area has a public relations value; that is to say, it serves the purpose of providing information that will ease tensions in the community. For example, a school

may conduct a survey of reading skills of sixth graders in order to demonstrate that the children have learned to read as well as those in communities where more traditional methods of teaching reading have been used. However, such attempts to build understanding in the community may involve the school in technical difficulties, for the pupils whose achievement is to be studied may be naturally more rapid or slower than those in the other communities with whom reading skills are to be compared.

Surveys of factual information designed to satisfy particular pressure groups also present dangers and may result in the modification of the curriculum in undesirable directions. For example, in one state the well-meaning local patriots have forced on the schools a curriculum that requires that the history of the state be studied every second year for the twelve years of public schooling. This has been the result of rumors of ignorance on the part of pupils of particular details in the history of the state, and no doubt a survey of the achievement of the pupils would show up some ignorance in this respect. However, it is extremely probable that no change in this situation would occur if the time devoted to state history were halved or doubled. It is probably similar to the situation revealed by Joseph Mayer Rice, who found near the turn of the century that it made little difference whether pupils spent one hour or five hours daily on spelling; in either case the achievement of the pupils was the same.

An ambitious attempt to study pupil attitudes by survey method is provided by the Purdue Public Opinion Panel, which conducts an opinion poll in cooperating high schools, mainly at the junior and senior levels. Schools that participate pay a very small fee to help cover expenses and in return obtain a report on the responses of their own students as well as on the responses of a wider sample of pupils. Surveys conducted in this manner are usually related to matters of rather widespread interest, such as the attitude of pupils toward various aspects of the curriculum or toward their parents. The results serve the purpose of stimulating thought rather than that of solving specific problems.

Much of the material collected about pupils by means of surveys is collected through the medium of paper-and-pencil devices, but information about pupil behavior can sometimes be collected by other means. It is possible, for example, to obtain records of the number

of books borrowed by each pupil from the library if a survey of reading is being made. The consumption of foods in the cafeteria may provide some evidence of eating habits in relation to health. Absentee rates are sometimes of considerable interest. Artistic products and other products of the pupils' hobbies may provide evidence of how leisure hours are spent. There is a wealth of objective pupil data that can be incorporated in a survey and that does not depend on the verbal responses of the pupils.

Data collected about the pupil is likely to have a relevance to the planning and development of education that data derived from the other two categories do not have. Its directness does not make it necessary to invoke questionable assumptions to justify its application to the solution of real problems.

Levels of Complexity in Surveys

Survey studies are mainly of the "what exists" type; that is to say, they are designed to determine the nature of an existing state of affairs. They may be considered to be research in that they result in the accumulation of a certain type of knowledge, not, in a sense, that their scientific status may be questioned. The reason for this is that scientific knowledge must consist of an organized body of generalizations that will permit the prediction of events which have not yet occurred and which are not predicted on the basis of what might be called common sense. The survey does not aspire to develop an organized body of scientific laws but provides information useful in the solution of local problems.

A genuine science of behavior in educational systems could never be based on surveys, but this should not necessarily deter the student from undertaking work in this area. Such surveys not only may solve problems of immediate importance but also may provide data that form the basis of research of a more fundamental type. As an introduction, it may be pointed out that surveys vary greatly in the level of complexity of the problems that they attempt to investigate. The simplest surveys attempt only frequency counts of events, while the more complex may seek to establish relationships among events.

The frequency-count type of survey. In this type of survey, the sole purpose is to determine the frequency of occurrence of a particular type of event or condition. The best-known surveys of this sort are those designed to determine the number of persons in a group

who expect to vote in a particular way at a future election. Schools may conduct surveys to determine how many children have received immunization shots of various kinds or how many teachers have M.A. degrees. Such surveys necessarily provide limited but often highly useful knowledge. However, they cannot be said to contribute in any way to an organized body of scientific knowledge about education.

The interrelationship-of-events type of survey. In many surveys, much more than a mere frequency count is sought; in addition, an attempt is made to find the interrelationship among events. Familiar studies of this type are those published by Dr. Gallup in which are shown tabulations of the voting preferences of adults split according to their socioeconomic level. Such surveys are usually based on some kind of theory concerning the interrelationship of events, and indeed should not be undertaken unless they are based on some fairly definite theory. It would be quite ridiculous to conduct a survey to discover the interrelationship of the physical characteristic of red-headedness and voting behavior. On the other hand, it would be quite reasonable to study the relationship between religious preference and party preference, for it is easy to see numerous reasons why there should be such a relationship. The study of such relationships is unlikely to provide direct evidence concerning causal relationships, and indeed they are extremely difficult to interpret. Although religious affiliation, for example, may be related to preference for one of the major political parties, it is clearly unreasonable to conclude that affiliation with religion *causes* the individual to vote for Republican candidates. Yet it is clear that such a relationship, if well established, must be a result of complex causal relationships that may produce both phenomena. The survey itself is unlikely to establish the nature of these complexities and is thus likely to leave the pollster with little except the bare facts, which form only a basis for speculation.

Surveys that approach experimental conditions. Under certain conditions, surveys may acquire many of the characteristics of experimental studies. For example, a researcher may be interested in the effect of hunger on drives and motives, and if he is fortunate, he may be able to obtain volunteers who will starve for varying lengths of time. He might seek evidence of the relative strength of food, water, sex, and other drives and motives by studying the fantasies and dreams of his subjects as starvation progressed, and

perhaps he might attempt the use of projective techniques for this purpose. The experimenter would also use a control group fed a normal diet, and comparisons would be made between the starved group and the control group. The starved and the fed groups would be carefully matched on the basis of relevant characteristics, or subjects would be assigned at random to one or the other group. In describing this experiment, it would be said that the experimenter *manipulated* starvation; indeed, in experimental studies it is customary to manipulate one of the major variables.

A rather similar purpose can be achieved through a survey technique. If the scientist could select groups already exposed to starvation by the natural course of circumstances and compare the fantasies, etc., of this group with those of another group that was well fed, he would be reproducing many of the features of the experiment already described. As a matter of fact, it has been possible to do this through the reports of persons who were exposed to extreme conditions of starvation in concentration camps in Germany during the war, and such studies have provided some evidence that as starvation proceeds all motives tend to become progressively more depressed except for those related to the obtaining of food.

There is, nevertheless, a major difference between the survey technique and the experimental technique. In the experiment described, a control group would have been included as well as an experimental group, and subjects would have been assigned to one or the other group in some way that could not systematically bias the results of the experiment. In contrast, in the survey technique in which a comparison of starved and fed persons is made, there is no assurance that these two groups did not differ systematically in some way that would affect the experimental results. Only when the assignment of persons to the starved and fed groups is under the control of the investigator—and this is true only in the case of the experimental procedure, only then is it possible to be moderately certain that some irrelevant influence is not systematically biasing the experimental results.

SURVEYS OF BEHAVIORAL PHENOMENA

The surveys of the conditions that constitute the environment of the school child, particularly the surveys of the physical conditions, do

not manifest the complexities that surveys of behavioral phenomena may involve. True, school surveys involve difficulties of their own, but much useful data may be obtained by straightforward objective techniques, which can be applied only rarely to surveys of behavioral phenomena. The transitoriness of most behavior that is directly observed, including verbal behavior, presents difficulties of observation and recording that a school building, a library, or a teachers' college transcript does not.

Surveys of behavioral phenomena within a school system may be concerned with the behavior of pupils, teachers, parents, superintendents, school boards, and other persons connected with the educational process. Such surveys may be concerned with verbal behavior such as the expression of opinions or desires, or nonverbal behavior such as whether parents do or do not spank their children. They may also be concerned with determining the distribution of relatively enduring traits such as intelligence, authoritarianism, schizophrenia, and a host of other attributes. Less enduring traits such as attitudes and interests may also be surveyed. Public opinion polls represent surveys at an even more transitory level. There are almost endless possibilities for surveying human behavior.

Desirable Characteristics of Behavioral Data Collected in Surveys

Little, if anything, is to be learned from a survey that seeks to collect a few items of information from a population that is selected merely because it happens to be at hand. While surveys of behavioral phenomena are limited in what they can produce because of the nature of the technique that they involve, they do not have to be as limited as this. In addition, they do not have to be just a process for collecting a large array of disconnected items of fact. The collection of masses of disconnected facts runs counter to the very first principle of survey design—namely, that the items of information gathered should be interrelated within a plan or framework.

It is of particular importance in most educational surveys to avoid phenomena that are transitory. This is of special importance in the behavioral area. Observations that refer to physical conditions do not usually refer to events that are of only transitory significance. When they do, it is so obvious to the researcher that it cannot slip by unnoticed. In conducting a school survey, it is evident that the

number of works of fiction in the library is a matter of more than transitory significance. True, it will change as time goes by, but only slowly. The square footage of floor space per pupil is also a matter of some permanence, despite changes in birth rate and in local sociological conditions. On the other hand, classroom temperature at a given hour on a particular day would probably be of only the most transitory significance and might provide little evidence of the quality of the physical conditions. Such a measure is quite clearly not one that could serve any useful purpose in a survey, and on this account it would be rejected immediately by any competent researcher.

In the case of behavior measures, it is much easier to overlook the fact that a measure may represent only an incidental phenomenon. For example, if a survey is made of the likes and dislikes of pupils for various school subjects, it is quite possible that the expressions of the pupils in this regard may be solely the product of the personalities of the teachers involved. The liking and disliking of junior high school students for particular school subjects may not represent in any way a fairly permanent structure in their personalities, though at higher levels (from what is known) they may well be considered to be so. The same is true in the planning of public opinion polls. It is useless to ask for opinions about matters concerning which the public has not formed a well-crystallized opinion. Opinions given about matters that are new, and for which there has not been sufficient time to think through the issues and form opinions on a solid foundation, are not generally appropriate matters for surveys. Opinions given on the spur of the moment are likely to be prompted by incidents that would have little weight after the subject had been given careful consideration. This is a matter that must be taken into account, not only in the survey type of study but also in studies involving instruments of the psychological inventory type. Interest tests commonly suffer from the defect of requiring the student to express his liking or disliking for an activity in which he has never engaged and in which he has never considered participating. It seems reasonable to hypothesize that preferences thus expressed may reflect only what are commonly called snap judgments.

The fact that the behavior elicited by the surveyor's questions refers to a relatively stable response should not be taken to mean

that this is any guaranty that the responses can be used for making inferences of the type that the investigator wishes to make. An example of this is shown by responses to a question that has been included quite frequently in public opinion polls. This question as asked is, "Do you favor labor unions?" (Sometimes it is stated in the form, "Do you believe in labor unions?") Responses to this question show, almost invariably, that even groups that are known on other bases to be antagonistic to labor unions give high percentages of "Yes" responses. There is every reason to believe that these responses are highly consistent even over a substantial period of time, and that they therefore represent enduring qualities rather than transitory phenomena. Despite this fact, these responses cannot be taken to represent underlying structures in the personality the assessment of which would permit the prediction of behavior in situations that involve labor unions. Indeed, the response seems to reflect a common stereotype of behavior and to reflect the fact that to be in favor of labor unions is to assume a socially desirable position. Nearly everybody is in favor of labor unions, just as nearly everybody considers himself a good citizen. For similar reasons, individuals like to call themselves liberals, and even those who might be more adequately described as conservatives may prefer to use the term "middle-of-the-road liberal." Such stereotyped responses yield little information, except possibly the frequency of occurrence of such stereotypes. However, it is extremely difficult to identify cases where such clichés are operating.

The Role of Theory in Conducting Behavioral Surveys

Surveys of behavioral phenomena need to be most carefully planned if they are to yield useful data. The information-gathering process should be based on some theory of the nature of the phenomenon that is being investigated. In most cases this is likely to be a fairly complicated matter, and it may require very extensive information if answers are to be found to important questions. A classic case of a behavioral survey based on well-developed concepts and calling for a large number of items of information is presented by the first Kinsey study of sexual behavior in the human male. In this study, several hundred items of information were collected about each male included in the sample. The purpose of obtaining such

extensive data was to be able to check and cross check a number of hypotheses concerning sexual behavior. The type of survey conducted by Kinsey should be contrasted with that conducted by Gallup and those whose data are published on a nation-wide basis through the newspapers. The latter type of survey collects a very limited amount of data about a matter of vital concern to the public. It is not designed for the purpose of developing understanding concerning the nature of voting behavior. Its usual purpose is to predict a particular event such as the outcome of an election, and although the pollsters have come to realize that some understanding of voting behavior may contribute to the accuracy of prediction, this is a secondary and minor goal. In contrast, Kinsey was not concerned with keeping a newspaper column going. His interest was a scientific one, and its goal was that of providing an organized account of certain aspects of human behavior, with some attempt to discover interrelationships.

The contrast that has been presented between the public opinion poll and the survey that has scientific purposes has had the purpose of pointing out to the student of educational research that polls published in daily newspapers are inappropriate models for the conduct of research. If educational research is to derive useful information from surveys, it is necessary that they be conducted at a proper level of complexity and based on well-thought-out concepts.

The theoretical framework used as a basis for a survey should, as far as possible, be stated as a theory in the way that has been outlined in earlier chapters. The limitation on this arises from the fact that surveys are commonly conducted in situations in which not much is known about the phenomena concerning which inquiry is made, and some of the data-collecting may be analogous to the grasping of a blind man in a new environment. The result of such grasping is the acquisition of information that will help him in developing a concept of the way in which his environment is arranged. The analogy does, of course, represent an extreme case, and the fact is that researchers, at least in their early years of development, would be most unwise to explore entirely unknown territory. They would be more likely to achieve positive results and to have a profitable learning experience if they were to conduct an inquiry into an area that had already been partially explored and in which earlier workers had already developed concepts for use in understanding.

Where limited surveys are conducted, despite the advice that has been given here to the contrary, the theory involved may be stated in the simplest terms and often may consist of no more than a single sentence. Perhaps a good beginning for the development of a conceptual framework is to recognize that a survey that accumulates data by asking questions and recording the answers is not just an easy process of obtaining information.

The matter of stating questions and obtaining answers is far from being the simple one that it is commonly believed to be. Perhaps all who have used this technique for collecting data during the past twenty years have shown a surprising naïveté in this respect, and we owe a debt to J.W. Getzels (1954) for calling our attention to our lack of sophistication.

Getzels has pointed out that, while numerous studies have been conducted to show that all kinds of conditions affect responses to visual patterns, no parallel research exists to describe conditions that affect responses in question-and-answer situations. He has borrowed a number of concepts from the field of perception to build a model that attempts to account for variations in the answers given to the same questions as the situation in which the question asked is changed. According to the model provided by Getzels, the asking of a question first produces an internal response, which is not verbalized. This immediate response is, in a sense, an answer to the question, and it is described as the personal hypothesis in this theory. Second, responses are made to various aspects of the situation in which the response occurs. This is referred to as the stage in which the demands of the situation are sized up in terms of the individual's personal adjustment to that situation. Third, the individual formulates a response that will facilitate his adjustment to the total situation.

According to this theory, there would be a tendency for the respondent to reply to a question in such a way that his answer reflects what he perceives to be an appropriate response to that situation. This is likely to involve answering in the way that he believes is expected of him.

The point here is that a theory to be used as a basis for a survey involving verbal responses must take into account the adjustments that are made between interviewer and interviewee. This complicates considerably the theory requirements for this type of work and recog-

nizes at the outset that data thus collected cannot be taken at face value.

At a minimum level of development, a theory must postulate the ways in which the phenomena can be measured. If various methods of measurements are possible, then there should be some statement of how these measures are related one to another. For example, if a survey is being made of parental disciplinary actions in the home, one might postulate that the parent's account of frequency of disciplinary action should be related to the child's account. If the two accounts are unrelated, then the data are quite probably worthless. One might also postulate that a more truthful answer might be given if a parent were asked to indicate what he thought would be the best way to handle a common behavior problem than if he were asked how he would handle his own child.

A theory on which a survey is based must usually postulate the relationships of the phenomena investigated to possible causes. In the case of the disciplinary behavior of parents, a survey would be a barren and dull affair if it stopped at the point of finding out how frequently parents inflicted different types of punishment on their children. The investigator would almost certainly want to know something about the characteristics that differentiated parents who punished frequently from those who punished rarely. To do this, it would be necessary to construct a theory of the determinants of punishing behavior. This theory would then form the basis of the inquiry.

A theory on which a survey is based must have something to say about the specificity or generality of the phenomenon that is to be surveyed. If the behavior of parents is being surveyed concerning the extent to which they attempt to exercise control over the behavior of their children, it is important to know whether the phenomenon is a general one; that is to say, it is important to know whether a parent who exercises extensive control in one area of child activity also tends to exercise control in other areas. If no information is available concerning this matter, then the survey must be conducted so that it samples the various areas of parental control and provides information concerning them. Unless such information is given, the results of the survey will have little meaning.

A basic question to answer in the planning of a behavioral survey

is, "What is the section of the population in which the phenomenon is to be found?" If the survey is to be concerned with an educational problem such as the existence of certain kinds of reading disabilities, the surveyor should know whether he is to be concerned with sixth or twelfth graders. The determination of this fact alone might have a desirable limiting influence on the scope of the investigation. Although it has been pointed out already that breadth is a most desirable characteristic of surveys, it is better to make an extended survey of reading disabilities of eight-year-olds than it is to collect less data about the entire population of school age. If the survey is to be strictly descriptive, the surveyor will proceed to determine what characteristics of reading skill are to be studied, in whom these characteristics can be most meaningfully established, and what are the sections of the population for which breakdowns are to be reported. It is probable that, if the reading skills of eight-year-olds are to be surveyed, the results would be reported for different socio-economic levels, by sex, and by school. The latter would represent three variables that previous research has already established to be associated with the development of reading skills. A survey that does not yield such breakdowns is a nebulous affair, providing an impressionistic type of picture that lacks any detail.

Types of Data Collected in Behavioral Surveys

The data collected in surveys may vary in the degree to which they represent directly the phenomena in which the surveyors are interested. The collection of data directly about behavior relevant for an educational survey is difficult to achieve. Examples of such data are found in many studies—as, for example, those in which pupils' food choices in the school cafeteria have been studied. This can be done by direct observation, and such a procedure is obviously much superior to that of asking the pupils what they will eat or have eaten. Spelling behavior may be surveyed by the inspection of samples of pupils' work. There is some evidence, incidentally, that the latter procedure provides information superior to that which can be derived from spelling achievement tests. Such data, however, represent only the most fragmentary records of a person's behavior in a particular area. It is possible, if not probable, that deplorable deficiencies manifested by pupils in their choice of food in the school cafeteria

may be justified in terms of their diet over an entire day. The sample of behavior, like most samples of directly recorded behavior, is too restricted and narrow to provide much useful information concerning the total eating habits of the pupils. Most surveys that attempt to obtain data directly without resorting to the collection of verbal reports suffer from this defect. This does not mean that observational techniques of the types previously discussed cannot yield valuable data and form the basis of worth-while research, for they can. The implication is only that they are not well adapted to research needs in those cases in which surveys are being conducted.

A second source of data for surveys about behavior is found in already existing records. Previous school grades and test records are examples of such data. Comparisons of the performance on tests administered at the same level but ten or twenty years apart represent a particularly interesting use of this technique. There is a very substantial body of data in most school systems about pupil personnel and teacher personnel, which may form the data of useful surveys. A word of caution should be injected here. Such data often include gross inaccuracies due to clerical errors made at the time when they were recorded. Further discussion of this matter will be presented later.

Since the collection of data about behavioral phenomena for the purposes of conducting surveys must involve mainly a question-and-answer procedure, it is necessary at this time to discuss, even though briefly, the matter of formulating questions.

The design of the questions to be used has come to be recognized as a matter of crucial importance in the planning of surveys. Suggestions for the design of questions can now be found in many sources. The reader is referred to sources such as Cantril's excellent book of over a decade ago (1947) and Parten's fine work (1950) on the subject. Some texts on educational measurement, such as that by the present writer (1955), also offer suggestions to follow in the design of questionnaires.

It would be inappropriate to attempt to discuss here at length the design of questions to be used in surveys, since this is a subject about which there is an extensive technical literature. Nevertheless, it is necessary to provide a brief orientation.

First, note that the design of effective questions is much more

than a matter of writing out clearly what one wants to ask. It is much more than a matter of the effective use of English. The question should be regarded as a stimulus to which there is a relatively stable response. If there is not a stable response or set of responses, then there is little point in asking the question, for responses to it will lack what is ordinarily termed reliability.

Second, the questions should be such that the responses are made to the questions themselves rather than to other aspects of the situation. For example, questions asked in a survey about the pay of teachers are likely to elicit a very high frequency of response to the effect that they are underpaid. Yet the same individuals who state that teachers are underpaid are likely to vote against increased state and city budgets that would make increases in their salaries possible. In answering the survey questions, these individuals are responding to the social pressures of the situation as much as to the question itself. Such responses are commonly referred to as stereotyped responses.

Third, the questions should draw responses that can be elicited with some uniformity. This is really saying that the procedure should have reliability and be reproducible. This is clearly not a sufficient condition to make a question useful, because questions to which there are stereotyped responses provide high consistency but little else.

Fourth, the questions must be such that inferences can be made from the responses to responses in other situations. This may not seem obvious when it is first considered, but it is precisely what is wrong with the question that elicits a stereotyped response—one given in the particular situation but presumably not given in other situations. If a survey is made to determine how the electorate will vote in a certain election, it is important to be able to assume that behavior in answering the survey questions will be related to behavior in the voting booth. If these two aspects of behavior are unrelated, the survey ceases to have any purpose. All questions asked in the survey method must be such that it is reasonable to assume some generality of the response. The reader should be reminded that the relationship between verbal behavior and other aspects of motor performance is complex, and that a simple one-to-one relationship can rarely be expected.

The trend in collecting behavioral information about populations of individuals has been to standardize the questions asked by the

interviewer. The opinion poll as it is commonly conducted represents a series of brief but standardized personal interviews, in which interviewers always ask the same questions and in which the responses of those interviewed are generally restricted to a few categories. It is possible to remove restrictions on the behavior of both the interviewer and the interviewee until the point is reached where a relatively free interview takes place, perhaps restricted only to the topic to be discussed. However, in surveys the difficulty of quantifying the data based on free interviews is such that they are not practical. The most that is likely to be done in departing from the situation in which both the question and the responses are restricted is the elimination of restrictions on the response. Such open-ended questions provide a range of responses that must ultimately be tabulated and codified.

The problems of collecting data by interview methods have been considered in detail in other parts of this book, and they cannot be considered further here. Despite the defects of interview techniques that have been noted, they must still be considered as the primary methods of collecting survey data about behavior.

The Direct-Mail Questionnaire Methods

The last twenty years has shown a transition in survey techniques from the use of questionnaires sent through the mail to quite elaborate interview techniques. It is unfortunate that the same trend has not been apparent in educational research, where direct-mail techniques are still commonly used and where even the research bodies of national organizations may rely upon such techniques in the development of their projects.

The central difficulty in all direct-mail techniques is that the percentage of returns is small. A questionnaire of some interest to the recipient may be expected to show only a 20 per cent return, even when conditions are favorable. If nonrespondents are then contacted a second and a third time, the return may be increased to 30 per cent. Only rarely does it reach the 40 per cent level. Attempts may then be made to contact personally the final group of nonrespondents, but if this is done, it might be as well to perform the entire operation by interview.

A considerable amount is now known about who does and who does not respond to mailed questionnaires. At one time it was felt by

many that it was largely a matter of chance whether a person did or did not respond to a questionnaire sent to him through the mail. If it arrived at a convenient time, he would respond to it; if it did not, he would not respond. Research has shown that this is not the case at all. A study by Wallace (1954) is particularly revealing in this respect.

In the Wallace study, four questionnaires were sent at intervals to the same group. Some failed to return any, while others returned as many as all four. The tendency was for a person who returned one questionnaire to return all four. In other words, there are those who by and large tend to return questionnaires, and there are those who do not. Insofar as these groups differ in relevant respects, the results achieved with direct-mail questionnaires are likely to be biased.

Wallace's study also throws some light on the general nature of the differences in the characteristics of those who return direct-mail questionnaires and those who do not. Of particular importance is the fact that those who return them show a marked tendency to have a college education, while those who do not have relatively less education. Since differences in education are related to a host of other variables, it is hard to find an area in which questions are not likely to elicit a different response from a well-educated group than from a less well-educated group. Thus the returns may be considered to be biased by an unknown quantity. It is of some interest to note that while the returnees and nonreturnees differ in education they differed little in the Wallace study in socioeconomic status. Wallace states that the safest rule in deciding whether or not to use direct-mail questionnaires is: Don't.

Checks on the Data-Collection Process

It is apparent from what has been said that the collection of verbal survey data is likely to leave the investigator with feelings of insecurity concerning the meaning of the information collected. It is usually recognized that empirical errors have been introduced into the data, and the size of these errors needs to be estimated if the results of the survey are to be interpretable. The experienced survey researcher will build into his study empirical checks that will provide him with information concerning the meaningfulness of his results.

Of primary importance are checks on the adequacy of the data-

collection process itself. The early organizers of surveys of consumer preference discovered that paid interviewers sometimes omitted the important detail of actually conducting interviews and adopted the short-cut process of filling out the interview schedules at home. Many ways have since been devised for checking on the honesty and accuracy of interviewers. One of these is to include on the interview schedule questions such that the distribution of responses in the population to be interviewed is known. Another common procedure is for interviewers to obtain the name and address of each interviewee and for a sample of these interviewees to be followed up independently.

Certain types of errors, known as response errors, are particularly difficult to estimate and control. Those who respond to interviewers are likely to show a tendency to overestimate those characteristics that are highly esteemed. Estimates of education and income are likely to be inflated unless they are preceded by questions that ask for such details as make subsequent falsification of answers hazardous. For example, if it is desired to determine a man's income, it may be as well to start by establishing his place of employment, his grade within the organization, and the base rate of pay for that grade, before direct questions pertaining to income are asked. The early questions in such a series provide data against which subsequent responses can be checked.

The internal checking of responses is probably the simplest and commonest way of testing the consistency of data. A person's given age can be verified against data such as the age of his eldest child and the age of the person when the child was born. The latter two questions can be separated from one another by other material and also separated from the original question on the topic. Sometimes a question on birth date can also be included as an additional check. It is possible for a respondent to be consistent in answering all of these items and to provide a completely false record, but the chances of this are small.

A type of check that has been used extensively for determining the trustworthiness of responses to personality test inventories has also been used in the conduct of opinion polls. The technique is simply that of asking a question to which the reply itself provides some estimate of the extent to which the respondent is providing trustworthy answers. If a person is asked, "Do you ever tell lies?" One

may suspect that he is not telling the truth if he answers "No!" There are few, if any, who never tell lies. It is possible to introduce a series of such questions, which may be used to provide a so-called validating score. Nevertheless, the technique is not without its pitfalls. The reader should take note of the fact that such questions, if they are answered honestly, usually place the respondent in a rather bad light, and for this reason the tendency to falsification may be much greater than would be the case with more innocuous items. The answers to such so-called validating items, or scores derived from them, cannot be unequivocally interpreted at this time.

If checks based on internal consistency are used, it is sometimes possible to determine the adequacy of these checks by separating from the population a group that may be expected to have an unusually high consistency of response. For example, if questions are asked school personnel about the behavior of individual pupils, one may expect principals to be less consistent than teachers, simply because principals have fewer opportunities to observe pupils. If a check showed that the principals were more familiar with the pupils than were teachers, the data would be open to suspicion.

The information given by internal checks provides evidence mainly of the internal consistency of the data. This is information related to reliability rather than to the trustworthiness of the data as a basis for inference. For this reason most opinion polls include checks whereby a part of the data collected are compared with equivalent data from some other source. A rich source of data in the educational field, which may be used for making many such comparisons, is that collected by the Purdue High School Opinion Poll. Other sources in the educational field are limited, and for this reason difficulties are encountered in the application of this method of checking the data. Hence a second and much less satisfactory method is commonly used; it involves the collection of observations additional to those ordinarily gathered in survey interviews. We are referring here to observations made by interviewers concerning the forthrightness or evasiveness of those interviewed and such behavior as may indicate that trustworthy responses are being given.

The Identification of the Sample to be Surveyed

In solving strictly local educational problems, surveys are commonly conducted to cover every member of the designated population.

In such cases we have no sampling problems, but there is also no population to which the results can be generalized. For example, in one state a survey was made of the attitude of members of the state legislature toward a particular educational proposal. Seventy per cent of the Republican members and 60 per cent of the Democrats were favorable toward the proposal. Can a test of significance be applied to determine whether the one party differed "significantly" from the other party in attitude toward this issue? The question is not a meaningful one so long as the total population has been polled, for a test of significance would attempt to answer the question as to what was the probability that such a difference existed in the total population if certain differences were found to exist in a sample. If the members of the state legislature can be considered to be a sample of, say, future and past legislatures, then a test of significance may be appropriate. However, we might have great difficulty in defining a population of which the members of the present state legislature could be considered to be a sample.

Sampling becomes a problem when it is desired to make a generalization from a sample of a specified population, either to other samples not yet drawn or to the population itself. The problem of sampling arises simply because it is desired to make a generalization.

There is an important distinction between a random sample and a representative sample. A *random sample* is one drawn in such a way that each member of the population has an equal chance of being included. Therefore, a random sample is one drawn without the guide of relevant variables in terms of which the sample should match the universe, and drawn in such a way that it is not systematically biased by the procedure in one direction or another. Of course one random sample will be expected to differ from another random sample; and, other things being equal, the larger the two random samples, the less will be the expected differences between their means on the characteristic sampled. On the other hand, in a *representative* or *stratified sample* cases are selected in such a way that the characteristics of the sample are similar in important respect to the characteristics of the universe. Thus in polling the public on their expected choices in forthcoming elections, it is common practice to select a sample so that it is representative of the entire voting public with respect to certain characteristics that are related to voting

behavior. For example, since voting behavior is related to income, it is important that the distribution of income in the sample be similar to the distribution of income in the universe. A similar control may be exercised over numerous related variables if the characteristics of the universe are known. This process of matching the sample to the universe permits greater validity of inference from the sample to the universe and from the sample to other variables than when the sample has been selected at random.

The researcher should avoid preparing an alphabetical list of names and then proceeding down the list until he has included a sufficient number in his sample. It is well known that some letters of the alphabet include more names from certain European groups than others, and this may produce bias in a sample selected on the basis of name alone.

In order to avoid the difficulties that are likely to arise from sampling an alphabetical list, a good plan is to select every fifth or sixth or tenth name, or whatever interval will yield the needed number while still distributing them over the entire list. Those who are absent on the day when the data are collected should be included as soon as possible thereafter. Substitutes should not be sought for these cases, since it is possible that absentees have relevant characteristics that the substitutes do not have. Also, there should be as little delay as possible in the testing of the absentees, for in any interviewing period the characteristic that is being measured may change. The researcher should also avoid the practice of depending on volunteers as a basis for selecting a sample, since volunteers are likely to be differently motivated from nonvolunteers. In the latter connection, the author can recall an instance where volunteers were compared with nonvolunteers on the Rorschach. The main differences between these two groups was that the volunteers came to the testing situations determined to reveal as much as possible of their inner life, while the nonvolunteers wanted to reveal as little as possible. Related influences may well affect scores from achievement tests.

Problems of sampling in educational surveys at the local level can often be completely avoided by including the entire universe of possible cases in the "sample." If a high school principal wishes to conduct a survey of the reading skills of the twelfth-grade pupils in his high school, it may be possible for him to include every one of

them in the survey. In most high schools this would be a practical matter, but in large city high schools where enrollments may be as great as ten thousand, it becomes quite unnecessary to test every pupil in the senior class in order to obtain the desired data. The decision concerning which of the pupils to include in the survey is greatly facilitated by the fact that there is available a complete list of the names of all cases in the universe from which the sample is to be drawn. When such a list is available, major difficulties associated with the problem of sampling are eliminated. Nevertheless, there are certain errors that the researcher should avoid.

Another common error in sampling school populations is to sample by seating position in an assembly hall. This would happen if the first four rows in an assembly were retained for a brief period to fill in a questionnaire while the others were dismissed at the end of the meeting. If the pupils were free to seat themselves in any way they wanted in an assembly, it is likely that those who chose the rear of the room would be different in many respects from those who chose the front of the room.

Difficult problems of sampling arise when it is not possible completely to identify the universe to be sampled. The typical public opinion poll, the results of which are published in newspapers and magazines, faces this difficulty. If a representative sample of the adult population is desired to determine public attitude with respect to some issue, there is no way of locating and identifying in advance of the poll the names of those to be questioned. This fact makes it difficult to insure that the sample is representative of the universe, or that it has not been selected in some way so that a systematic bias has been introduced. Fortunately, most populations that the educator may wish to sample have been inventoried for him, as is the case with populations of pupils and parents. However, when he wishes to extend his inquiry outside of these groups, he runs into the difficulty of identifying the population that is to be sampled and therefore has difficulty in recognizing the cases that are to constitute the sample.

The problem that this situation presents is a difficult one, and we must pause here to consider some of the historical solutions and their weaknesses. The oldest method of all was that of merely avoiding the problem and including in the sample just those cases that could be easily located or that volunteered. The disastrous *Literary Digest*

straw ballot on the outcome of the 1936 Presidential election was run on this kind of a basis. Millions were included in the sample—indeed, it was one of the largest ever included in a ballot—but circumstances beyond the control of the investigator caused the sample to be biased, and inferences made from the sample about the universe were unjustified, as the results clearly showed.

The public opinion pollsters who followed immediately after this debacle recognized that some control needed to be exercised over the characteristics of the sample. It was also recognized that those who volunteer opinion, as in the *Literary Digest* poll, may be so different from the total electorate that it may not be possible to select from these cases a sample that could be considered representative of the total for the purpose at hand. A similar bias would occur if a principal polled those attending the P.T.A. and assumed this to be a sample of all parents of all the children in the school.

The trend in the 1930's was to distribute the group included in public opinion samples geographically and to instruct interviewers, who were also distributed geographically, to obtain interviews with certain specific percentages of individuals in each economic structure, each race, each sex, etc. For a long time this appeared to be a satisfactory procedure, and more by good fortune than by anything else, the predictions of the outcomes of national elections enjoyed a decade of apparent accuracy. However, it eventually became clear that the procedure of letting interviewers select those to be interviewed results in a bias in the sample thus selected. Interviewers tend to select interviewees who are rather like themselves. Thus, if the interviewers belong to an upper socioeconomic group, they are likely to include too few individuals in the lower income brackets. Interviewer bias is now a well-known and well-established phenomenon.

Some adjustment also had to be made for the fact that only some of those polled voted, and those who failed to vote represented a biased sample of the electorate. In addition, some last-minute changes in opinions were likely to occur, and not necessarily such that the proportions in each voting category remained unchanged. These adjustments were usually made on the basis of judgment based on past experience, and for many years this was a successful process. However, it is well known that in the 1948 Presidential election a failure occurred in the making of these adjustments, and the result-

ing predictions were notoriously inaccurate. Since that time, those concerned with the conducting of surveys have become more cautious in the making of such adjustments and have also worked on improved methods of obtaining samples for this purpose.

One of these newer approaches is referred to as *area sampling.* In this technique, highly detailed maps of the regions to be sampled are used and the area is systematically sampled. If, for example, a particular small area is to be included in the sample, then all persons living within that identified small area are included in the sampling. The sample thus selected is largely independent of the whims, likes, and dislikes of the persons collecting the data. Nevertheless, the method is not as simple as it seems, and difficulties are encountered in tracking down the persons identified. There are also definitional problems. If a person has residence in a particular locality, it does not necessarily mean that he lives there, and decisions have to be made about such matters.

In some localities, samples may be identified in advance of the process of collecting data by obtaining complete lists of residents. Here again, the procedure appears to be highly satisfactory on the surface, but difficulties arise in its application. Lists of residents are often inaccurate because of faulty methods of collecting the data on which they are based. An even greater source of difficulty arises from the time lag between the collection of data for making lists of residents, its publication, and its subsequent use for survey purposes. In many areas, a lag of only a year may render such lists quite unsuitable for identifying any kind of sample. On the other hand, other areas may show a high population stability.

Special problems arise when follow-up surveys of school graduates are made. Studies of these groups are commonly undertaken in order to determine the successes and failures of the graduates so that the program of the school may be improved. In these follow-up studies, it is easy to obtain a complete list of the names of the population to be studied. The difficulty arises in locating these individuals. Young groups are particularly mobile, but fortunately their parents represent a population much more stable in terms of home address. Many can be contacted through their parents' homes, but in some localities it may not be possible to do this, since entire families may move to different parts of the country. If the survey is a follow-up of college

graduates, then the alumni organization may be of great value in locating individuals. Classmates may also be consulted to determine addresses of cases that cannot otherwise be located. The investigator must be resigned to the fact that in the educational follow-up survey it is likely that a substantial fraction of the cases to be included in the sample will not be located despite intensive efforts on the part of the investigator.

Sometimes in conducting follow-up studies, one can check on the extent to which the sample collected is representative of the total population included in the study. For example, if the school files still retain the scholastic records of the population, it is possible to determine the extent to which the sample is representative of the universe that is sampled with respect to scholastic achievement. This would be an important fact if scholastic aptitude could be considered to be related to the phenomenon under study. If, in this case, a serious discrepancy existed between the sample and the universe, questions might be raised concerning the validity of inferences from the sample to the universe.

In the development of follow-up studies of graduates of particular educational institutions, a special problem occurs in the choice of specific classes to be sampled. Suppose that the purpose of the study is to obtain information concerning the adequacy of the curriculum for preparing pupils for the future. Perhaps it is considered that a sample of sufficient size can be obtained by including 50 per cent of the students from a single graduating class. In such a case, the investigator probably might become concerned with the problem of selecting a typical class, so that the results could be generalized to other graduating classes. The investigator would soon find that this task presented some real difficulties. He would note that recent graduating classes would be inappropriate for the purposes of the study, since the students would have been out of school for too short a time to permit inadequacies in their school experience to reflect themselves. He might therefore consider including in the study a class from several years earlier. This plan might look good to him until it was pointed out that these earlier classes had fed most of their male students into the armed forces, since the students graduated at a time when the military establishment was being built up to meet a war emergency. These students faced situations that later students did not face. Still earlier

students also had unique experiences of their own, since they were thrust on a labor market where unemployment faced many. The investigator would soon be forced to accept the fact that no graduating class could be considered as typical, because education takes place in a changing world.

However, one presumes that there must be at least some uniformity running through the varied environmental conditions that present themselves to graduating classes. At least one may expect that most members of all graduating classes ultimately face the problems associated with earning a living. Insofar as the graduating classes show such a common core of experience, it is feasible to make inferences from data derived from one class to other classes.

Sometimes the researcher may arrive at the conclusion that no class and no sample will provide the information needed. No survey is a panacea. The investigator nevertheless will exercise great caution in making the inference that the results derived from one graduating class can be generalized to other classes. Even though all graduates face the problems of earning a living, economic conditions vary from class to class. One class may graduate in times of depression, another in times of prosperity, and the resulting problems and difficulties may be entirely different. Indeed, one might even be forced to the unsatisfactory conclusion that the curriculum needed to prepare pupils for prosperity might be different from that necessary to prepare pupils for depression conditions.

Some Misuses of Survey Methods

The survey as it exists today within the framework of educational research finds its greatest misapplication in the local study of the type that educational institutions usually undertake for the purpose of justifying their existence. The difficulty with such surveys is that they are designed to answer questions that really cannot be answered at the present time by means of any data that one can conceive of collecting. For example, the writer has watched the development of a study designed to provide an evaluation of selected aspects of teacher education. Some of the questions that those conducting the study proposed to answer were truly answerable, such as, "What has been the effect of state aid on the program?" Such a question can be answered mainly by consulting the budget office and by determining just how

the money has been spent. However, an answer is likely to add but little to what is already known through the channels of common gossip. Most of the answerable questions posed by the committee running the survey were of this character. However, it was the questions that were much more difficult to answer, if they could be answered at all, that dealt with the problems of central importance to this survey. For example, among the questions were, "Does the institution graduate students who can teach successfully?" and, "What are the weaknesses and strengths of the teacher education program?" Such questions, unfortunately, cannot be definitively answered by any means at present available.

Now in the case of the particular survey, those who asked the questions soon realized that there was no satisfactory way of answering them. They saw that there was no way of determining whether the teachers produced by the program could teach successfully. Thus an alternative question was formulated: "Are the principals who employ the graduates of the college satisfied with these graduates as teachers?" In order to answer this question, a questionnaire was sent to the principal of each school employing one graduate or more. This questionnaire asked for information concerning the extent to which the graduates were satisfactory as teachers. As one might well expect, those principals who did reply rated the teachers trained in the particular institution almost always as satisfactory, or very satisfactory. This, of course, meant absolutely nothing. A person likely to make a derogatory report would probably make no report at all. Thus such "data," if one may excuse this misuse of the term, were quite valueless for answering any significant questions that might be asked about the teacher education program.

A Final Word on Survey Methods

Although it has been said that survey methods result in knowledge of a low order—that is to say, knowledge that permits little generalization—this should not be taken to mean that surveys are easy to undertake. This is far from the case. Surveys should be undertaken only by those who are aware of the difficulties and who have some mastery of the techniques. The survey should not be thought of as a simple method of research, for it is as difficult to execute as any other type.

SCHOOL SURVEYS

The development of methods for undertaking surveys has been intimately related to the development of accreditation procedures. This is hardly surprising, since the accrediting associations represent the major enterprises that engage in school survey work, although professors of education, educational consultants, state and local superintendents, and others also engage in surveys of schools to varying degrees. Accrediting associations, however, have had to enter into the business of conducting school surveys with a certain amount of system to their methods, since they are open to public criticism.

The development of methods for evaluating schools by inspection has been closely linked with school and college relationships. In this process, the early stage was marked by the admission of students to college by examination, a procedure that is notorious for its tendency to standardize the high school curriculum in terms of college admission requirements. The recognition of the fact that cooperation between secondary schools and colleges was a necessary prerequisite to the development of secondary and higher education led to the founding of the New England Association of Colleges and Secondary Schools in 1884, the Association of Colleges and Secondary Schools of the Middle States and Maryland in 1887, and the North Central Association of Schools and Colleges in 1895. The latter grew out of the Michigan School Master's Club, which held an annual meeting at which secondary school and college teachers discussed problems of mutual interest. The members of this organization believed that it might be profitable to bring together teachers from a wider area, and thus the North Central Association of Schools and Colleges was formed. However, the early college and school associations were primarily devices for bringing together individuals to discuss and solve common educational problems, and it was not until 1901 that any activities were undertaken to develop a system of accreditation. In the latter year the Commission on Accredited Schools was established to investigate this matter. The Commision was given broad authority to set up standards for high school courses that would be accepted for credit by colleges, and to set up standards for accreditation in general.

Parallel with the work of the North Central Association was that of various state departments of education, which were concerned with the problem of accreditation from the standpoint of setting up minimum standards at which all schools should aim. Since that time, the function of accrediting secondary schools has become one shared by large state universities and state departments of education, which in many states have performed this function cooperatively and at times interchangeably.

While the initial interest in this area was in the accreditation of specific courses, the procedure soon broadened out to include a multitude of matters, such as the length of the school year, the number and length of the periods given each week in each subject, the training and experience of the faculty, the size and scope of the library and other physical facilities, and other matters too numerous to list here.

A study by McVey (1942) pointed out that the North Central Association of Schools and Colleges has been a powerful influence in the establishment of standards for secondary schools by the various states. He goes on to point out that this is partly a result of the fact that influential members of state departments of education have often been members of this association and have attended meetings where standards have been established.

The basic problem in the development of procedures for accrediting schools is to establish a basis on which schools are to be judged. Clearly it is insufficient to turn loose an observer in a school. Observation must be restricted to certain aspects of the school and its program that are considered of central importance to the effectiveness of the program. The essential characteristics observed during the accreditation procedure are referred to as *evaluative criteria.* The remainder of this section attempts to describe the general nature of the evaluative criteria that have been used in the accreditation of schools and colleges; that is to say, it describes the types of schedules that have been developed for guiding observers who are sent out to obtain information about schools by inspectional procedures.

The Accreditation of Secondary Schools

The most comprehensive attempt to draw up a guide for the evaluation of schools was an outcome of the Cooperative Study of Secondary School Standards, which was first organized in 1933 by the

representatives of six major regional accrediting associations. The results of this study appear in numerous scattered publications. The purposes were the following:

1. To determine the characteristics of a good secondary school.
2. To find practical means and methods to evaluate the effectiveness of a school in terms of its objectives.
3. To determine the means and processes by which a good school develops into a better one.
4. To derive ways by which regional associations could stimulate and assist secondary schools to continue growth.

The same study also provided a series of schedules for evaluating secondary schools in the following areas:

Agriculture
Art
Business Education
English
Foreign Languages
Health and Safety
Home Economics
Industrial Arts
Industrial Vocational Educational Core Program
Mathematics
Music
Physical Education for Boys
Physical Education for Girls
Science
Social Studies
Progress of Studies
Pupil Activity Projects
Library Service
Guidance Services
School Plant
School Staff and Administration

Furthermore, the volume on evaluative criteria derived from the study includes a schedule for evaluating individual members of the faculty. Various statistical and graphical devices for summarizing the data thus collected are also included. In addition, schedules are provided for determining the extent to which the school is meeting the educational needs of youth and for determining the nature of the child population served by the school.

Each one of the schedules for the evaluation of work in specific subject-matter fields organizes the evaluation into the following areas:

1. Organization. This covers such matters as how the curriculum is developed, whether there is continuity in the organization of studies in the area, etc.

2. Nature of offerings. This category explains itself fairly adequately, but it does include such matters as whether the courses provide opportunity for student responsibility and leadership.
3. Physical facilities. This covers such matters as furniture, visual aids, and general classroom conditions.
4. Direction of learning, divided into the four areas given below:
   a. Instructional staff. This covers preparation, background, organization, etc.
   b. Instructional activities.
   c. Instructional materials.
   d. Methods of evaluation.
5. Outcomes. This covers assessments of what students have learned in the program, though few hints are offered as to how the assessments are to be made.
6. Special characteristics of the program in the area.

Under each one of these areas a check list is provided against which a mark or other symbol is entered according to the following system:

++	Provision of the condition is made exclusively.
+	Provision of the condition is made to some extent.
X	Provision of the condition is very limited.
M	Provision of the condition is missing but needed.
N	Provision of the condition is not desirable or does not apply.

On the basis of all the evidence in any one area studied, an over-all evaluation is made of the effectiveness or worthwhileness of that aspect of the operation. These evaluations are summarized on a five-point scale, on which the points are as follows:

Scale Value	*Interpretation*
5.	Excellent: the provisions or conditions are extensive and are functioning excellently.
4.	Very good: (a) the provisions or conditions are extensive and are functioning well; or (b) the provisions or conditions are moderately extensive but are functioning excellently.

Scale Value	*Interpretation*
3.	Good: the conditions or provisions are moderately extensive but are functioning well.
2.	Fair: (a) the provisions or conditions are moderately extensive but are functioning poorly; or (b) the provisions or conditions are limited in extent but are functioning well.
1.	Poor: the provisions or conditions are limited in extent and are functioning poorly.
M.	Missing: the provisions or conditions are missing and needed; if present they would make a contribution to the educational needs of the youth in this community.
N.	Does not apply: the provisions or conditions are missing but do not apply or are not desirable for the youth of this school or this community.

The items listed under each heading of each evaluation sheet vary in specificity. Some are highly specific, and many ask whether staff members have had specific types of experience, as in the case of the item that asks whether home economics teachers have had actual work experience in this field. Some are so general that it seems almost impossible to determine whether the condition exists or does not exist. For example, it may be almost impossible to answer in terms of the categories provided whether the program of a school "is based upon an analysis of the educational needs of youth," for it is not clear whether it is to be based on systematic investigation. Also, it is not clear what is meant by "the educational needs of youth," for are these to be needs already experienced, or needs in terms of the problems they will face later in life? The term "need" is one with a multitude of meanings. As another illustration of the same difficulty, one may wonder how it is possible to determine whether a program "encourages enlargement and enrichment of the pupil's scope of interests."

The lists of evaluative criteria display no pretensions of being comprehensive, and indeed spaces are provided on the schedule for the addition of items that are relevant to the specific situation in which

the evaluations are made but may not apply outside of those situations.

Any criticism of the schedules prepared in the cooperative study of school standards must take into account the purposes for which they were prepared and the background of thinking on which they were based. A superficial examination of the schedules reveals that they seem to bear some resemblance to orthodox psychological and educational measuring instruments but that they do not meet customary standards of acceptability. This criticism is not entirely fair, even though it may be pointed out that the end result of the use of the schedules is a single numerical rating based on a series of evaluations of a number of important elements in the situation. In addition, it may be pointed out that the ratings thus arrived at are produced by a highly subjective process and cannot be appraised in terms of norms because no norms are available. Finally, the measurement expert might point out that no evidence is given concerning the reliability of the assessments provided by the schedules, nor is there any evidence concerning the validity of these assessments. These criticisms are not entirely logical, for the following reasons:

First, the history of school inspection and accreditation during the last fifty years has illustrated a trend away from the use of quantitative data and a return to qualitative standards. Therefore the schedules that represent a recent stage of thinking in this area do not represent a series of measuring devices to be used in a standard way; rather are they guides to the thinking of the person who is undertaking the evaluation. They present a series of topics that *may* be given consideration in the total assessment procedure, and it is recognized that some of these may be irrelevant in some situations and that some relevant ones may have been excluded from the list. Some guide to thought is better than for each assessor to be entirely his own guide.

Second, numerical norms of the type provided by most publishers of achievement tests would be largely meaningless in the assessment of secondary schools, since different schools must be assessed by different standards. The curriculum provided by a large secondary school serving an industrial population must differ in some ways from that of a small school serving an agricultural community. The failure of schools of the latter type to meet the needs of an agricul-

tural population is the most common criticism of professional visitors to them, and yet this is an entirely different criticism from that directed against schools in industrial communities.

On the other hand, the criticism concerning the lack of evidence of the reliability or validity of the recorded assessments cannot be passed off lightly. If individuals cannot show substantial agreement with themselves or with others in the entries made on the schedule with respect to a specific school, then the schedules and the records made on them have no value. Evidence of reliability would be fairly easy to obtain, and the only real excuse for its lack is the large amount of money that such an undertaking would probably involve. Evidence of the validity of the end products of the schedules must also be produced. No escape from this problem can be offered by any argument that the schedules are valid by definition, for the condition of validity by definition does not exist.

The Accreditation of Colleges

The need for developing a system for accrediting colleges arose from a different source than that which stimulated the development of machinery for accrediting secondary schools. Zook and Haggerty (1935, 1936), who have reviewed this matter, conclude that the movement for the establishment of standards for the accrediting of colleges arose from a need for exercising some social control over higher education, which expanded so vastly during the first half of the present century. These authorities point out that, although there are many ways in which some public control may be exercised over higher education, control by accrediting associations offers the advantage of freedom from political pressure and controversy. Accrediting associations may honestly raise standards of education without being accused of political intrigue or influence.

It may be noted that the problems of establishing criteria and of establishing standards are quite distinct. Two different accrediting agencies may use similar criteria, and yet because their standards differ they may vary in the percentage of institutions inspected that they accredit. Strictly speaking, the establishment of criteria should precede the establishment of standards.

The development of procedures for accrediting institutions of higher education has had a history covering over fifty years, and contributions to these procedures have been made by numerous indi-

viduals, many accrediting associations, divisions of the federal government, state departments of education, the American Council on Education, and other organizations and individuals. However, an overview of the total situation would indicate that many of the major developments in the procedures have come from the North Central Association of Schools and Colleges, which has sponsored some of the few systematic studies in this general area.

In 1934, after many years of deliberation, this Association published a manual (1934) to be used in the accrediting of colleges and a series of schedules on which the data relevant to the accrediting procedure were to be recorded. In addition, a series of monographs published during the years 1935 and 1936 provided extensive data on the use of the criteria described in the manual, and even went so far as to provide some normative data for some of the measures used in the accrediting procedure.

The manual provides criteria for evaluating each one of the following aspects of a college:

Faculty	Student personnel services
Curriculum	Administration
Instruction	Finance
Library	Plant
Induction of students	Institution's study of its problems
Intercollegiate athletics	

Each one of the areas to be evaluated is broken down into elements. Consider the matter of evaluating the faculty. This is first broken down into the areas of (1) faculty competence, (2) faculty organization, and (3) conditions of faculty service. For the first of these three, criteria are listed for determining the degree of competence. Some of these criteria are:

Percentage of total staff holding an earned doctor's degree.

Average number of years of graduate study of the staff.

Average number of years of experience in teaching and administration in institutions of higher education.

Number of scholarly books and monographs produced per staff member.

Number of memberships in national learned societies per staff member.

Number of places on national programs per staff member.

The criteria listed in the original manual were tried out over a period of several years, and in 1941 a revision of the criteria was published. In the revised manual a substantial amount of normative data is provided to assist in the interpretation of data collected during the accreditation process. The normative data are based on the institutions of higher learning accredited by the Association. As an illustration of these data, it may be noted, with reference to the percentage of the staff holding an earned doctor's degree, average values are given for junior colleges, teachers' colleges, liberal arts colleges, and universities, and in each case separate data are given for publicly and privately controlled institutions. In another table, data are presented showing the average number of books and monographs and articles published per faculty member in each of these types of institutions.

The inadequacy of the evaluative criteria both at the secondary school and at the college level is well recognized by the Association in its reports, which frequently emphasize the fact that the program of a school must be evaluated as a whole. It is also recognized that there seems little possibility at the present time that qualitative criteria can be replaced by quantitative criteria, and that fundamentally the process of accreditation must depend on subjective judgment. Frequent cautions are given that the criteria outlined should be considered more as a guide to thinking than as a series of basic dimensions along which assessment must be made. The emphasis on caution gives recognition to the fact that the process of assessment in this area is still in the earliest stages of development.

Criticisms of Evaluative Criteria Used in Accreditation

A number of important criticisms of the evaluative criteria discussed in this chapter must be considered, but these criticisms must be reviewed in the light of the fact that this type of measurement is relatively new, about as recent as the development of tests of intelligence. Nevertheless, these forms of measurements have not had as much concentrated thought bestowed on them by so many people as have tests of intelligence. Work on accreditation procedures has been largely the spare-time activity of relatively few individuals.

First, it may be noted that there is no experimental basis for the evaluative criteria commonly used in the inspection of schools and

colleges. There is general agreement that the main ultimate criterion of the effectiveness of an educational program is the extent to which it produces desirable changes in the pupils. Evaluative criteria for use in accreditation are based on the judgments of educators that certain characteristics of a school do have an effect on the extent to which the objectives of learning are achieved. It is assumed, for example, that it makes a difference in the amount of learning accomplished whether a faculty of a secondary school does or does not have professional training in the courses provided by departments in education. As far as the present writer knows, there is no evidence that teachers with professional training of this type are more effective than those who do not have training of this type. One may suppose that it would make a difference, and all teacher-training is based on this assumption, but many assumptions made by educators in the past have been shown to be unjustifiable on the basis of scientific experimentation.

Second, there is a problem well recognized even by those who have developed the evaluative criteria for the North Central Association. It is that the attempt to achieve rigorous standards of measurement may prevent the assessment of the outstanding characteristics of an institution. A secondary school may be performing a first-rate job even though its faculty has had limited training and the plant is poor, but the desire to do a professional job may overcome deficiencies of formal training, and ingenuity may make up for deficiencies in the plant.

Third, normative data may have relatively little value since they do not set minimum standards but only show how one institution compares with others. On the norms provided, one institution appears to be low and apparently inadequate because others are higher on the scale, although the fact may be that all the institutions are inadequate.

Fourth, the normative material provided was developed during a period of great educational change, which included times of oversupply and undersupply of teachers. These changes would make it difficult, if not impossible, to use norms of the types provided, because by the standards provided, institutions would show great changes even from year to year.

Fifth, the system of evaluative criteria does not take into account

the fact that single items may be crucial. A school that has a program quite unrelated to the needs of its students should not be accredited even if it is adequate on all of the other dimensions listed. A rural school that fails to take into account the fact that most of the pupils will eventually enter agricultural pursuits is inadequate, even if it achieves high scores on other variables.

One may assume that eventually all of these criticisms will be met after careful studies have been made of the extent to which the various factors are related to the degree to which the objectives of learning are achieved. Before this can be done, it will be necessary to develop valid measures of a great number of outcomes and to measure the outcomes of teaching under a variety of conditions. The problem is complicated by the fact that different institutions have different objectives, and consequently the achievements of the pupils in one place may not be comparable with those of another.

An Overview of Accreditation Procedures

Survey procedures for assessing the effectiveness of schools and colleges in achieving particular objectives must be considered as relatively crude methods of appraisal. They are all based on numerous assumptions, some of which are open to question. Although the validity of these procedures may be questioned, the process of inspection has certain intrinsic values that may justify it regardless of validity. First, accreditation and inspection procedures are becoming more and more a service; that is to say, they are designed to help schools and colleges improve themselves rather than to act as a threat. Thus accrediting agencies now often provide the services of special consultants to help schools with special problems. For example, the University of Michigan functions as the accrediting agency for the secondary school of Michigan, and as a part of its function it provides consultants in a great many different areas.

Second, accreditation procedures encourage schools to examine themselves. This is always a healthy process, and a well-organized accreditation agency can perform a valuable function by encouraging schools to do this.

Although accrediting agencies may use the most primitive methods of assessment and measurement, their usefulness cannot be questioned when their power is exercised with wisdom.

Summary

1. Survey research methods represent research on educational problems at a rather simple level undertaken mainly to solve problems of local significance only.

2. Surveys conducted in educational research are commonly undertaken as efforts to determine the nature of the physical conditions related to education. Sometimes surveys are made of the behavior of teachers or pupils. A further type of survey attempts to establish the achievements of pupils.

3. Surveys may merely enumerate the frequency of occurrence of some type of event, or they may study the interrelationship among events.

4. Surveys may attempt to undertake studies that could be undertaken by experimental methods, but they do not provide the same certainty of knowledge that experimental procedures might provide.

5. Surveys of behavioral phenomena should not represent a mere effort to collect a set of unrelated facts. The information gathered should be interrelated within a plan or framework.

6. Surveys should avoid obtaining information about transitory behavioral phenomena. They should also avoid questions likely to produce a mere stereotyped response that the respondent feels to be appropriate for the occasion.

7. A survey should be based on a theory of the nature of the phenomena that are to be surveyed. Since surveys are often conducted in areas where relatively little is known, it is often difficult to develop an adequate theoretical basis. All surveys that involve a question-and-answer approach should be considered as studies involving a complex social interaction between a questioner and a respondent. The theory should specify the general nature of the phenomena to be investigated, the methods through which aspects of them can be measured, the conditions that produce them, and the population in whom the phenomena are to be found.

8. The direct observation of behavior in naturally occurring situations has limitations as a survey technique. It usually represents a highly selected sample of the total daily behavior of the individual. Surveys conducted through the administration of tests or through an examination of pupil products have had a long history of utility.

9. Problems and difficulties involved in the design of questions for surveys have been extensively explored by research workers, and the person who undertakes a survey involving the asking of questions should be familiar with what is known about the preparation of such materials.

10. Direct-mail questionnaires should be avoided unless no other method is available for obtaining the desired information. Those who return questionnaires delivered through the mail tend to be a more educated group than those who do not.

11. In any survey, checks should be built into the data-collection process itself. The main type of check used is an examination of the data for internal consistency.

12. Since sample surveys are designed to obtain information from a sample that can be applied to a universe, it is most important that the universe to which the results are to apply should be specified and that the method of obtaining the sample should be an appropriate one.

13. The research worker who conducts a survey should be sure that the resulting data will be meaningful. Too often the results of surveys provide biased information to support some person's prejudices.

14. The development of school survey techniques has been intimately connected with the development of accreditation procedures. In this connection, the North Central Association of Schools and Colleges has played a leading role.

15. The characteristics of a school that are observed during the accreditation procedure are referred to as evaluative criteria.

16. The Cooperative Study of Secondary School Standards was a comprehensive attempt to provide a guide for the evaluation of schools. This guide provided a system of rating scales through which the observations made concerning a school could be quantified.

17. The trend over the last half-century has been away from the use of quantitative standards, which have come in for serious criticism. The main difficulty in using such standards is that different schools have to be assessed along different dimensions.

18. Similar attempts have been made to provide quantitative criteria for the evaluation of colleges.

19. The numerous criticisms that have been leveled against present systems of evaluative criteria indicate that there exists here a fruitful field for research. There is a real social need for a continuing program of research in the area. The assumptions underlying the use of current accreditation procedures need to be investigated.

Some Problems for the Student

1. As a result of a school survey, an attempt was made to obtain a score indicating the qualifications of each teacher. This score was derived by adding together the number of years of teaching experience, the number of semester hours of professional training beyond the bachelor's

degree, and a rating of "teacher effectiveness" provided by the principal. What assumptions are made in adopting this procedure?

2. It has been planned to determine by means of survey methods the number of those trained for teaching who remain in the profession after five and after ten years, and the reasons for attrition. Develop a theory that accounts for losses to the profession over the ten-year period. Suggest procedures that might be used for determining the rate of loss, and identify the assumptions on which the procedures are based.

3. A superintendent of a large city school system is aware of the existence of extremely low morale among the teachers. He realizes that there is widespread mistrust of the administration of the school system, but he wants to obtain information that will provide him with a basis for making decisions that will right the situation. He can obtain a substantial sum of money to conduct a survey but is afraid to proceed lest his intentions be mistaken. What approaches might be adopted in conducting a survey so that the full cooperation of the teachers would probably be maintained?

4. Outline a procedure for conducting a survey of the arithmetic skills of the fifth grade children in a large city school system serving a community having a population of half a million. Assume that only 25 per cent of the pupils are to be sampled. What meaningful norms or standards could be used in the interpretation of the data, which are to be used for public relations as well as normative purposes?

5. A superintendent has requested his research department to interview all teachers leaving the system in order to obtain information that might help in retaining a greater number in the future. The request is specifically for an interview, and a questionnaire cannot be used. Plan a structured interview to collect information about this problem, basing it on a theory concerning the reasons why teachers leave. Suggest internal checks that might be introduced in order to estimate the validity of the data obtained. Suggest other sources of data that might be used to check those obtained in the interview.

Prediction Studies | 11

Research on Problems of Prediction

Survey research at its simplest level attempts to determine the nature of a particular universe of events. "What is the state of affairs that exists?" is the type of question that surveys most commonly attempt to answer. Inferences are sometimes made from the sample that is examined to the universe that is sampled, and occasionally predictions are made of future events, such as election results; but the major purpose of surveys as they are conducted in educational research is not the prediction of events in the future. However, a great many educational research studies are carried out with the primary object of developing methods of making predictions. This type of study and problems of its execution must be considered in this chapter.

Prediction studies within the domain of educational research may be sociological, economic, or psychological. Attempts may be made to predict pupil enrollments at some future date. Predictions may also be made of the future supply and demand of teachers, and of funds to be available for teachers' salaries from direct taxation. Or study

methods may be developed for forecasting the success or failure of pupils in different curricula. Sometimes attempts are made to provide predictions over a relatively long period of time. Such studies are illustrated by those that attempt to develop methods of predicting college success from tests given in junior high school. These studies are concerned with problems of the greatest importance, for the assignment of pupils to a proper curriculum in high school must depend on the ability of school personnel to predict how the pupil's talents can best be used at a later time in his career.

The Pseudo Science of Predicting Something from Anything

A word must be said about the type of educational study that involves predicting *something* from *anything*. Usually both the *something* and the *anything* are rather vague. Many such studies begin with the graduate student's dissatisfaction with current procedures for predicting scholastic success in some field of study in which he is interested. Such a student may have been a high school teacher of accounting, greatly concerned with the fact that a rather large fraction of students who enter accounting courses fail to achieve satisfactory grades. He may feel there is a need for building a test that will eliminate those applicants who are almost certain to fail. Various tests have been tried out, but none has proved to be particularly useful. This teacher, therefore, decides to collect a number of new tests and administer them to students of accounting, in the hope that one will turn out to be a good predictor of grades. This might be called a shotgun approach, and it has disadvantages with which the graduate student should be familiar.

First, it is a departure from the type of scientific methodology that has yielded so much in the past and represents a return to a much more primitive method of achieving knowledge. It is a return to the kind of prescientific technique practiced by the medieval physician, who tried whatever herbs and techniques he had at his disposal in the hope that something would be found to help the patient. Occasionally this approach worked and the patient was cured, and in this way there accumulated a considerable amount of unconnected items of information that had their uses in the primitive practice of medicine. Such scraps of lore did not make medicine a science. Neither will large numbers of correlations between test scores and measures

of performance in handling life's daily problems of work and play constitute a science of behavior. Only when these apparently disconnected facts are integrated into a system is there any hope that they may form the rudiments of a science.

Second, even if a correlation exists between a test and the *something* it is desired to predict, there is always a real possibility that the correlation may be due to some irrelevant aspect of the *something*. For example, one might find that ratings of personal attractiveness of female college students were correlated with grades in college. One might certainly suspect that this correlation was generated by the fact that male college professors might have a tendency to overestimate the academic achievement of outstandingly attractive college women. Such a hypothesis would be much more reasonable than to suppose that personal attractiveness has a genuine relationship to academic achievement.

Third, since it is a hit-or-miss procedure, it is necessary to include a great many potential predictors—unless, of course, a theory is available that permits the more accurate predictors to be selected in advance. Many studies of the predictive value of brief biographical items of information have been carried out by administering several hundred such items to groups whose behavior it was desired to predict and then selecting the items that had the greatest predictive value. Such procedures are laborious, require extensive statistical treatment of the data, and are costly. They are most appropriate where useful results are to be achieved rapidly regardless of cost.

John Dewey (1910) has elegantly compared the relative merits of the empirical and the scientific methods of prediction in the following statement, which was written nearly half a century ago:

> While many empirical conclusions are, roughly speaking, correct; while they are exact enough to be of great help in practical life; while the presages of a weather-wise sailor or hunter may be more accurate, within a restricted range, than those of the scientist who relies solely on scientific observations; while, indeed, empirical observations and records furnish the raw or crude material of scientific knowledge, yet the empirical method affords no way of discriminating between right and wrong conclusions.*

* From *How We Think*, by John Dewey, copyright © 1910 D.C. Heath and Company. Reprinted by permission.

Statistical Empiricism

A few would raise their voices to defend the viewpoint that, since the laws of behavior are statistical laws, statistical methods should be the primary basis for discovering such laws. This statement means that the ultimate purpose of educational research is to derive a set of statistical relationships between conditions and the consequent behavior of the individual. The end product of the behavioral sciences, then, could be represented by something like a dictionary listing sets of conditions (both internal and external) and the probability of occurrence of each of various consequent behaviors. Perhaps the end product of the behavioral sciences may be such a handbook, but this product would be quite unsatisfactory to most with a scientific turn of mind. Perhaps the advanced stage of knowledge reflected by such a handbook may have appeal in contrast to our present ignorance, but it is easier to see it in its true light if we consider the possibility of developing a similar product in other fields where it is not so difficult to be objective. Consider, for example, the problem of weather forecasting. At one time, weather forecasts for a particular locality were made by keeping records of *what followed what* in the sequence of weather conditions. Thus a high southeast wind in a particular locality might be taken to indicate rain, since rain followed more frequently than anything else on the tail of such a wind. Nobody knew why this was so, but the probability of the occurrence was well established; but so long as nobody knew, there was no way of improving the accuracy of predictions. However, such a system of forecasting has been abandoned, because its accuracy could never be improved beyond that permitted by the data previously collected. The present system, which has replaced the old statistical system, is based on a knowledge of how weather conditions are produced. It is based largely on air-mass analysis and thermodynamics, and permits much more accurate predictions than the older statistical methods. This does not mean, of course, that mathematical methods are not used for making predictions today, for they are. However, their function is to use data in accordance with some complex theory of weather prediction. Modern weather forecasting has in fact become highly mathematical, and introduces the help of electronic computers in order that complex mathematical functions may be computed at a relatively rapid speed.

Despite what has been said here, there are two ways in which one may properly speak of statistical laws. First of all, the laws of physics are, in a sense, statistical laws. The statement that the pressure exerted by a gas on the walls of the vessel in which it is contained is uniform at all points assumes that atoms of the gas moving at different velocities are moving in all directions in roughly equal numbers, so that the force exerted by these atoms of the container is uniformly distributed over the wall. If this were not so, the pressure on one part of the wall would be greater than on another part of the wall, and one could not talk about uniform pressure of the gas in the container. Since there are large numbers of atoms or molecules of the gas involved, the probability is extremely small that all or most of them will happen to bounce off one small area of the vessel wall at one time and thus raise the pressure at that point and lower it at other points. In fact, this probability is so small that it is disregarded in the formulation of the law. This is one meaning of the term "statistical law"; but when it is said in educational research that all laws of behavior are statistical, a somewhat different concept is usually involved.

When reference is made to *statistical laws* in educational research or in other research in the behavioral sciences, an entirely different concept is usually being considered. It is the present writer's impression that this concept signifies a tendency toward regularity in the sequence of events. Such tendencies may vary from complete regularity to regularities just above what one would expect on the basis of chance. If these so-called statistical laws were referred to as tendencies toward regularity in events, any respectability that is derived from the use of the word "statistical" would be eliminated.

Tendencies toward regularity may vary from those in which there is no knowledge concerning the laws of the observed tendency to those in which there is fairly complete knowledge. Mere tendency to regularity, or what appears on superficial examination to be a tendency to regularity, is often the point at which the scientist starts his work, and at this starting point one cannot say that there is anything that at all resembles a scientific law. Once the scientist has established the variables functionally related to the tendency to regularity, he has acquired useful knowledge that goes beyond that provided by a mere statistical relationship.

Empiricism and Research on Problems of Educational Prediction

Research on problems of predicting educational achievement has not usually been scientific in the sense in which the term has been used in this volume. Inevitably this has been so, for the urgent need for making accurate educational predictions has prompted those concerned with the problem to grasp whatever facts were available. In addition, in the partial solution of urgent problems that are complex in character, it is often much more feasible to try out a large number of possible solutions and see which will work rather than it is to develop a program of research along systematic and scientific lines. At least three types of empirical procedures have been adopted in this setting, and the merits of each need to be considered.

Method I. The miniature-situation approach. This is a procedure for developing methods of scientific prediction that really involves no research at all, but simply requires the educator to reproduce a miniature and abbreviated situation in which the individual can be given, so to speak, a trial run. The experimenter hypothesizes that performance in the miniature situation will reflect quality of performance in the situation in which it is desired to predict behavior. Thus, in the development of algebra prognosis tests, an attempt has been made to introduce into the test situation some of the learning activities that the pupil will have to face in his first course in algebra. Language prognosis tests use a similar technique. One such test measures the ability of the student to learn a small amount of Esperanto. It has been shown that the ability to learn small amounts of this artificial tongue is related to the ability to learn large amounts of other languages.

This technique is generally a successful one. The major condition that may mitigate its use is that which occurs when learning in the early stages of an activity involves different abilities from learning in a later stage. Such changes in the determinants of behavior as learning progresses have been shown to occur in certain instances, but these changes have not been particularly striking and probably are not sufficient to prevent the use of a miniature learning situation for selecting those most likely to succeed. However this may be, activity directed toward the development of such a technique for a particular purpose cannot be said to make a contribution to scientific knowledge.

The product is a technique that in no way adds to available organized knowledge.

From the point of view of developing guidance practices, the miniature learning situation does not result in a product that fits well into current procedures. It is clearly quite impractical for the guidance counselor to administer as many miniature learning situations as there are situations in which one may desire to predict behavior. The guidance worker needs a short and comprehensive battery of tests that overlap as little as possible. A battery of miniature learning tests would show substantial overlap, with resulting inefficiency in the testing procedure. Guidance batteries that are currently widely used do not include the miniature learning situation type of test.

Method II. The hit-or-miss approach. This method has already been discussed, and it is briefly mentioned here in order to contrast it with the other methods. This approach to the problem of prediction involves the administration of a wide range of instruments in the hope that one will be found that predicts successfully. This statement is a little exaggerated, in that the investigator is unlikely to try out just *any* instrument, but rather will he select those that appear to have at least some remote connection with the phenomenon to be predicted. The technique finds support in the fact that it has had a long and fairly successful history of application. A strong point in its favor is that many a time an unpromising variable has turned out to be the best predictor. Once this has occurred, it is nearly always possible to find a good reason why it should be so. On the negative side, there are also several points to be noted.

The method involves a great amount of work, both on the part of those administering the tests and on the part of those taking them. There are real questions as to whether anyone is justified in occupying so much of another's time. *Careful thinking through of the problem might result in the tryout of a much more limited battery of instruments, with less time lost by all.* This gain must be balanced against any loss that may result from unlikely variables turning out to be good predictors.

In addition, the variables likely to be selected are those that have some superficial relationship to the phenomenon that is to be predicted. If an analysis of the prediction problem is made in terms of current psychological knowledge, it is probable that only a few likely

predictor variables will appear. but these may not have any relationship obvious to the layman to the predicted variable.

Method III. The scientist's approach to the problem of prediction. A third method. which is advocated in this book, involves the development of a theory concerning the nature of the phenomena to be predicted, and, on the basis of this theory, the derivation of methods hypothesized to predict. The writer at one time attempted to develop a theory that could be used as a basis for prediction studies in the achievement area. It was built around the concept that the predictor variables in prediction studies were intervening variables and that they had certain specified relationships to one another. The writer does not hold any particular brief for the theory he developed, and urges the student to develop his own miniature system.

Up to the time of writing, so little has been done to develop prediction studies on the basis of the third method discussed here that it is difficult to discuss the problems that it presents. The primary difficulty most certainly lies in the theory-construction phase itself.

The Time Factor in Prediction

In experimental studies, the experimenter begins by specifying a certain set of conditions existing at the beginning of the experimental period. Then certain experimental variables are manipulated for a given time, and the resulting changes in initial conditions are assessed. Thus substantial control is exerted during the entire time during which data are being collected. In contrast, in the type of prediction studies considered here, there are long periods during which occur events that are relevant to the outcome of the study but during which no control is exercised over existing conditions. This lack of control during long periods of time makes the results of such studies highly tenuous. The longer the interval between the prediction and the event to be predicted, the less are the chances of making a successful prediction.

Conditions Necessary for Effective Prediction

Many prediction studies end in failure that could have been avoided if the researcher had considered the problem carefully in advance. In many such cases, a careful consideration of the problem in the first place would have led to the conclusion that the prediction was not a

feasible one. The discussion that follows is designed to help the student determine what is likely to be and what is not likely to be a feasible variable to predict.

First, in order that a phenomenon may be predictable from a given point in time, it is necessary that the determinants of that phenomenon exist in an identifiable form at that time. If one accepts the principle of universal causation, then it follows that the conditions existing at any moment must include all the conditions necessary for accounting for all subsequent events. However, these determinants may exist within such a multiplicity of events that they may not be identifiable by any means at present conceivable. When such is the case, it is simply not feasible to attempt to undertake prediction.

This discussion may perhaps be abstract to the person who has not spent long years struggling with problems of predicting behavior at some future date, so it is desirable to expand on the matter by way of a concrete example. Consider the problem of predicting the number of teachers who will resign from a large school system during each year for the next ten years. This is no trivial problem, since the long-term training of teachers requires that candidates be trained to replace those resigning from the system as well as to take other positions that will have to be filled over the course of the years. Large numbers of resignations may leave gaps that cannot easily be filled unless there has been long-term planning.

The time when it is necessary to make the prediction of whether a given group of teachers will or will not resign is four or five years before resignations actually take place, since this is the time required for recruiting and training new teachers. This is a considerable span of years over which to make predictions but many educational forecasts are made with a useful degree of accuracy over this period.

An immediate suggestion about how the prediction should be made is that data be obtained from the past and applied to the future. It certainly would be possible to obtain data on resignations over a long previous period, say twenty years, and to work out an average resignation rate. On the surface, this may appear to be a good method, until the data are closely examined and the discovery is made that most of the resignations occurred during a rather short period during the war. The reason for this was that teachers then were offered wages in industry far above those that could be obtained within the educational system. If this condition did not recur within a com-

parable period of time, it would not be expected that the previous resignation rate would be comparable to the future resignation rate. However, even if such a period of high wages for ex-teachers did occur, school districts in the future might be willing to offer teachers a bonus or other financial incentive to stay on in the system, in which case the resignation rate might be held at a low and constant level. Insofar as the resignation rate depends upon economic conditions and international tensions, it is not predictable by any technique at present available—at least economists and political scientists have not yet succeeded in predicting such events and conditions.

An alternative approach can be taken if the the problem is redefined. In place of stating the problem as that of predicting the percentage of teachers who will resign in a given year, it may be redefined as that of identifying those who are most likely to resist the temptation to resign. Stability could be given to a teaching body if it included only those who are likely to stay with the system indefinitely. It seems reasonable to assume that the personal characteristics of those who remain might be different from those who resign. One might suspect that those who stayed would have a deeper interest in teaching and a more favorable attitude toward the activities it involves than those who would leave for economic reasons. One might perhaps hypothesize that those who stayed might tend to be less ambitious, and perhaps less intelligent, than those who did not. Conceivably a study could be designed to discover ways of identifying teachers who would not resign for economic reasons. At least some of the necessary conditions exist for practical predictions to be possible if the problem is stated in this way. However, it must be pointed out that the usefulness of a study of this kind might well be questioned. It would be hard to imagine an acceptable teacher-selection procedure that would permit the rejection of those who did not present characteristics making for long years of service. Indeed, such a procedure might well be criticized on the grounds that it eliminated some of the ablest teachers and those who might provide leadership for the system. However, it must be pointed out that the problem of creating a stable body of teachers should be attacked realistically by making economic adjustments, for economic conditions are clearly a major determinant of resignations, and attempts to solve the problem by selection would not attack at its roots.

From what has been said, it is clear that, for a phenomenon to be

predictable, the determinants must exist in some well-identified and measurable form at the time when the prediction is made. If partial predictions are to be accepted, and they must be because perfect predictions cannot be made, then only partial determinants need exist in an identifiable form. Since the determinants of political events remain largely obscure at this time—for we have had little success in predicting events in the social and economic sphere—prediction studies in the behavioral sciences must direct their energies toward the use of individual characteristics as variables. Many successful procedures for making predictions have been evolved through studies based on this type of approach. An excellent example is found in studies in the prediction of college success. As a result of these studies over the last quarter of a century, useful tests have been developed for identifying students who will later withdraw from college because of poor grades. This prediction can be made because the differences in grades are determined mainly by conditions within the individual. Such of these conditions as are related to differences in ability can be measured, but those related to differences in motivation cannot be adequately measured at the present time. The reader should note that although predictions in this area are quite accurate in terms of the level of precision one may expect in the behavioral sciences, they nevertheless leave much to be desired. The reasons for the rather large imperfections in our predictions must stem to a great degree from the existence of external conditions that affect grades. For example, such matters as whether the student is assigned to an instructor who has a personality compatible with his own, whether illness does or does not strike him, whether he falls in love or does not fall in love—all these and many others are unpredictable circumstances that introduce error into our predictions.

Another condition that must be established before a prediction study is undertaken is that the phenomenon to be predicted must be homogeneous in its causes, that is to say that it always has the same causes. An example of a condition that it has not been possible to predict because of the multiplicity of possible determinants is delinquency. It is obviously most desirable to predict which children are most likely to become delinquent, so that the clinical psychologist and social worker can get to work to prevent this from happening. The difficulty is that there are many major determinants of delin-

quency. Some delinquency is a product of lack of intellectual insight into what is happening. Other causes are the effects of associates, the home background, and various pathological psychological conditions, to mention but a few. Under these conditions, there is no single effective way of identifying the potential delinquent. This problem is discussed at greater length later in this chapter.

An additional important condition for prediction is that the condition to be predicted must represent a well-defined phenomenon, and, if possible, that it represent a measurable variable. A much discussed variable such as teacher effectiveness does not meet the necessary standards of clarity. On the other hand, if specific and well-defined aspects of teacher effectiveness are used in prediction studies, then there is danger that the researcher may be able to predict only the trivial. The discovery of a significant and well-defined variable to forecast is often the major difficulty in the development of a prediction study.

Research that is designed to evaluate the effectiveness of counseling frequently suffers from the fact that the condition to be predicted cannot be described in terms of a single variable. Although we may talk in generalities and point to *adjustment* as the condition to be predicted, there are many ways in which a person may adjust, and these cannot be compared to one another easily, if at all. In the face of this difficulty, many quite ridiculous criteria of the success of counseling have been evolved. For example, in one study the success of the counseling procedure was evaluated in terms of whether the counselee returned for more. A somewhat better solution might be to classify those who come for counseling into groups in terms of the type of adjustment to be made or the problem to be solved. Within any one group, it may be possible to distribute success at making the desired adjustment along a single and meaningful scale. Partitioning of groups into relatively homogeneous subgroups is often a solution to the problem of simplifying the conditions of prediction to the point where they are manageable.

Prediction Studies of Behavior as Studies of Enduring Traits

From what has been said, it is hardly surprising that most prediction studies are studies of aptitudes of types that are known to be relatively enduring. Alternatively, they may be studies of biographical

factors that are presumed to have an enduring effect on behavior. In any case, it is assumed that the uncontrolled events intervening between the time of prediction and the time of occurrence of the behavior it is desired to predict do not affect the magnitude or character of these traits. This is justified to a considerable extent, for it is known that the rank order of a group of children on an intellectual ability, such as that measured by a vocabulary test, does not change appreciably over a period of time that may be as long as a year or more. This property of long-term stability has resulted in the extended development and widespread use of such tests, for their predictive value is partly possible because of their stability over time. Studies of aptitudes have revealed a relatively small number of such traits in the intellectual field in the aptitude area that it seems profitable to measure. These are the variables that are commonly measured in aptitude batteries; beyond these there does not seem to have been much success in the measurement of highly stable intellectual variables.

As to nonintellectual attributes, commonly referred to as personality traits, little success has been achieved in using them for prediction purposes. This may be a result of their instability. There is, of course, considerable evidence that the amount of many social traits that a person manifests depends to a considerable extent on the situation in which he is placed. At a cocktail party he may be a warm and genial character, quite the reverse of what he is at the office. Insofar as traits vary with the situation, satisfactory conditions for long-term prediction do not exist unless the situation as to which predictions are to be made can be carefully described.

Difficulties in prediction also exist when the trait to be measured and to be used for predictive purposes is capable of extensive modification through experience. One may suppose that many social traits show progressive learning through the years of schooling. Unless the rate of learning is known, it is not possible to predict future behaviors from these traits. The rate of learning is almost certainly likely to be an unknown.

Finally, consideration must be given to the problem of using biographical data for the making of predictions. The use of such data is based mainly on the assumption that the exposure of the child to certain environmental conditions results in the development of

particular attributes that later become determinants of behavior. Difficulties in the use of such data arise because of the problem of identifying just what happened in the individual's past. There is little difficulty in determining *what he himself thinks happened*, but this may be quite different from what actually happened. Also, what *he thinks happened* will probably change from time to time, while what actually happened will not change. For this reason, among others, the predictive value of biographical events as they are reported has been found to be small.

The Availability of Appropriate Conditions

In the case of many prediction studies that students suggest, the difficulty of carrying them out lies in the unavailability of the conditions necessary for executing them. This is true of most of the studies of teacher effectiveness that are proposed. The misfortune is that so many of these are pursued under conditions that do not permit the production of meaningful results.

One could fill a volume describing the prediction studies that have been attempted, only for the researcher to discover that the conditions necessary for making the study did not exist or could not be found. Most studies related to the long-term predictive value of tests used in guidance are of this character. Only rarely is it possible to follow up those who have been tested at an earlier date. Studies of the prediction of leadership qualities from test scores also rarely can be followed through to a successful conclusion, if only for the fact that leadership is not a single behavioral dimension but a conglomerate of perhaps unrelated characteristics.

Studies designed to predict teacher effectiveness present some of the best examples of attempts to forecast phenomena under conditions that do not permit prediction. Such studies usually are based upon the assumption that an observer can spend a limited time in a classroom, perhaps an hour or two, and on the basis of what he sees can make a valid judgment of the effectiveness of the teacher. This, in turn, assumes that teacher effectiveness is a single dimension along which all teachers can be measured and compared. Both of these assumptions are nothing short of nonsense. The realities and complexities of the teaching situation are such that neither of them is in any way acceptable. One of the few definitive statements that one

can make about the classroom situation is that it presents phenomena involving a great number of variables. Undoubtedly, observer characteristics enter into the selection of items from this complexity and result in many of the peculiar properties of ratings of teacher effectiveness. This, however, is not the only problem encountered in studies of teacher effectiveness. A large number of other conditions must also exist before such studies can be made profitably. In order to bring to the attention of the reader the range of circumstances that must exist for a study of this kind to be successful, it is worth reviewing here one of the better-conceived studies in this area.

The study referred to here is one by Morsh, Burgess, and Smith (1955), who were concerned with the extent to which student ratings of instructors could be used to predict the extent to which the objectives of a particular course were achieved. Unlike most of their predecessors, these investigators were able to find a situation in which this prediction problem could be studied. The situation was that presented by a group of 106 instructors who were *all teaching the same course* in hydraulics in an Air Force installation. It was also possible in this study to obtain the cooperation of two successive classes for each instructor, each class consisting of about fifteen students. Students were assigned to instructors at random, but, as one might expect, these groups differed considerably in their ability to learn the particular type of technical subject matter. Students were given both a pretest and a posttest of the subject matter taught in the course, and these tests were carefully tried out on an independent group of students. On the basis of the data thus collected, the tests were revised and improved in order to increase their internal consistency. It was assumed in this study, and the assumption was justified, that the effectiveness of the instructor could be measured in terms of the extent to which the objective of the course was achieved. The latter was the acquisition of subject matter, and it could be measured in terms of the achievement tests. On the basis of the student's previous grades and measures of his academic aptitude, a gain score from pretest to posttest was predicted for each student. However, the pretest was made extremely easy, and the gain score was not simply the difference between the pretest and the posttest scores, but the numerical value of the posttest score corrected for differences in initial knowledge of the subject matter as indicated by

the pretest. The posttest scores were also corrected for differences in the learning ability of the students as measured by another aptitude test and by previous performance in school. Thus the gain score was the posttest score corrected as far as was possible for differences in the learning ability of the students.

Unlike their predecessors, Morsh et al. were able to demonstrate that the average gain score shown by the class of a particular instructor was something more than a transitory phenomenon. Successive classes given by the same group of instructors showed similar gain scores; that is to say, an instructor who produced a high gain in one class tended to produce a high gain in a subsequent class, and vice versa. Thus it was established that differences between instructors produced differences in gains in knowledge that were consistent from class to class. As far as the writer knows, the latter item of important information has not been positively established in any previous study. This does not mean, of course, that similar consistent gains would be found in the classes of other instructors whose teaching was in a different subject-matter field or directed toward different objectives. As a matter of fact, the writer is aware of at least one other study where the gain scores showed no consistency from class to class, and where its value as a variable to predict therefore rested in the shadow of doubt.

Morsh et al., therefore, clearly established that they were concerned with something more than a transitory phenomenon in the use of gain scores, before they went on to study some of the conditions associated with high or low gain scores in particular classes. These investigators were able to establish that certain types of ratings of the instructor made by the student could be used to predict the corrected gain scores. They were not able to establish relationships between particular aspects of instructor behavior and corrected gain scores. This seems to indicate that relevant aspects of teacher behavior that promote learning were not incorporated in the study. For the most part such aspects of teacher behavior have not yet been identified.

The Morsh et al. study represents a good beginning in the area, but only a good beginning. Advancement will be impeded by the fact that few researchers are likely to have available to them situations in which a large number of instructors are working toward identical

goals and with students assigned to their classes at random. Even if such a situation presented itself, it is doubtful whether most researchers would have the financial support necessary for preparing testing materials in quantity or for the employment of trained observers to obtain data concerning the behavior of the teachers in the classes. However, even if a favorable situation for studying problems of predicting teacher effectiveness existed, there is no guaranty that techniques for measuring aspects of teacher behavior are sufficiently advanced to permit extensive developments of knowledge. It is quite possible that investigations may not be able to proceed beyond the groundwork laid by Morsh and his associates.

The point stressed here is that the existence of a situation favorable to the making of a particular study may not insure success in advancing knowledge. It may be recalled as an example from another science that the existence of high-powered microscopes did not permit the study of microorganisms. Although these microscopes had sufficient power, they could not be used for examining microorganisms until the additional technique of staining these organisms had been developed. Most complex problems—and most educational problems are complex—become amenable to study only after a multiplicity of techniques have been developed for handling various aspects of them.

Morsh et al. were able to take advantage of a unique situation that permitted the undertaking of a study that probably could not have been undertaken in most other educational situations. The latter do not ordinarily provide particularly favorable situations for the conduct of educational studies. In most cases, it requires all the ingenuity that the researcher can muster to adapt a proposed inquiry to an available research situation and to adapt an educational situation to the purpose of an inquiry, so that the study can be undertaken in a way that yields meaningful results.

This study has been discussed in some detail in order to bring to the attention of the reader the range of conditions that sometimes must exist before a meaningful prediction study can be made. The student should list carefully the conditions that he must find in order for him to carry through a prediction study in which he is interested, and then check to see whether such conditions actually exist in the facility where the study is to be undertaken. If the necessary conditions do not exist, the study should be abandoned. Too commonly

in the past, researchers have proceeded with their prediction studies despite the fact that the results could not possibly be meaningful.

Fractionating Populations to Increase Accuracy of Predictions

A number of interesting cases have been found in which it has not been possible to make predictions for an entire group, but in which predictions could be made within a section of that group. For example, it has been found in studies of achievement motivation that in some situations this variable shows little relationship to performance when an entire group is involved. On the other hand, when it is possible to separate from the total group those who see the task to be performed as a challenge, then within this small group a marked relationship exists between achievement motivation and performance. This is not surprising, since achievement motivation can hardly be expected to operate in situations in which the individual does not feel a need to do his best.

In almost every area of educational research, one can think of situations in which it is necessary to partition a population of events in order to establish relationships. Where relationships are to be found between the qualifications of teachers and the characteristics of the curriculum, one would expect different relationships in urban schools than in rural schools. Sometimes it may be necessary to separate boys from girls in order to make a meaningful prediction. Sometimes it may be necessary to separate cultural groups. In other cases, relationships may apply to only certain types of economic conditions and not to others. A careful thinking through of most studies is likely to reveal the possibility that some of the relationships expected are more likely to occur in certain sections of the population than in others. It is of considerable interest to determine whether such hypotheses are sound.

Clustering of Variables to Increase Accuracy of Predictions

It happens frequently in educational research that a large number of variables are included as potential predictors of a particular phenomenon. These predictors may show irregular but low correlations with the variable it is desired to predict. It would be possible, of course, to compute a combination of the variables that will best predict the particular independent variable. If this procedure were

followed, and if the researcher were concerned with many predictions, he would be likely to find that a combination that maximized the prediction would provide what appeared to be an accurate prediction, but that when the same combination of best predictors was applied to a new sample, the prediction would shrink substantially. This is the well-known phenomenon of *shrinkage,* and it has dealt a fatal blow to many studies that were promising on the surface.

A second approach to the problem of building up predictions does not suffer from this hazard. It involves first the clustering together of those predictor variables that belong together in terms of their intercorrelations. This can be accomplished by means of factor analysis or by the related method of cluster analysis. Variables that cluster together are then combined in some way. Such composite variables may generally be expected to have the merit of having higher reliability than the relatively low-reliability elements of which they are composed.

In the clustering of such variables, a group combined should be constituted of elements that belong together, not only statistically but also according to a rationale. Unless this is done, any prediction made from the cluster is unlikely to contribute systematically to knowledge; rather is it likely to represent only an odd but perhaps useful relationship.

Just as variables within the predictor group may be clustered and then combined in the hope of improving the accuracy with which predictions may be made, so too may groups of independent variables be clustered. For example, an investigator concerned with the prediction of teacher behavior might have observed a group of teachers for the frequency with which they performed various acts, such as raising their voices, threatening to punish, offering rewards, asking for suggestions, encouraging a pupil to pursue a matter further, offering help, etc. The investigator would probably find that only the poorest predictions could be made of the extent to which a teacher manifested any of these categories of behavior. However, it is quite likely that a correlational analysis would show that some of these behaviors tended to cluster together. It would certainly be expected that all behaviors representing expressions of hostility would represent a cluster of correlated measures of behavior. When measures of all of these behaviors are added to form a measure that might be described as the tendency to manifest hostility—from what has been

learned about such a variable from other sources—one might expect that this characteristic of teacher behavior might be reasonably predictable from test scores.

An example may now be given of a case in which both the dependent and independent variables in a study consisted of composites. In this study, a series of tests of creative ability was administered in order to predict the creative aspects of public speaking. The test scores were combined into composites on the basis of a previous factor analysis, and the composites had considerably greater reliability than the original scores. In order to increase the possibility of obtaining the greatest amount of predictability, the measures to be predicted were combined into similar groupings. Thus all measures related to high-level originality were combined together, both among the predictor variables and among those to be predicted. The resulting composites might be expected to be related. It should be noted that this method does not have some of the disadvantages of the multiple-regression method of combining variables for maximizing predictions, in that the relationship is unlikely to shrink when it is tried out on a new sample. The problem of shrinkage is considered later in this chapter.

Clinical Versus Statistical Prediction—A Problem in the Validity of the Direct Observation of Behavior

In recent times, there has been considerable controversy concerning the relative merits of clinical predictions and so-called actuarial predictions. What is meant here by a *clinical prediction* is a judgment arrived at by a psychologist after considering a certain body of data. An *actuarial prediction* is made by combining quantitative data to derive a score, which is used to make a prediction. Clinical psychologists have generally maintained that it is possible to make more accurate predictions through the exercise of clinical judgment than could be made by the statistical treatment of data alone—at least insofar as it is treated by the methods at present in common use. The problem is an important one in the current connection, because it implies that the data-processing method of the researcher is inferior to that of the machine.

Various approaches have been taken to the study of this problem. One has been to compare the actuarial prediction with the prediction of the clinician made on the basis of the same test scores.

Meehl (1954) has reviewed the studies up to 1954 in which the accuracy of predictions made by clinicians using test scores have been compared with the results achieved by statisticians using objective methods. The results seem to vary considerably from one study to another, depending on the nature of the condition to be predicted. In no clear-cut cases did the clinicians predict more accurately than the psychometricians. One suspects that, where the psychometrician has a well-developed procedure for predicting a particular type of event or condition, he will do better than the clinician, but if he does not have such a procedure, then the clinician may possibly do better.

Just what can be concluded from the comparison of the statistician's predictions and the clinician's predictions made from the same data is hard to understand. It would indeed be immensely surprising if a clinician could improve upon a testing and statistical procedure that had been developed and refined over the years for making a specific type of prediction.

Conceivably the clinician is better than the statistician in certain situations, but the statistician may be more accurate in making other types of predictions. If, for example, it were desired to predict what the writer is most likely to do next Sunday, test scores would be a very poor basis for making a prediction. However, what he is likely to do can easily be predicted from a knowledge of his habits. It would seem that whenever the behavior to be predicted is based upon well-established individual regularities of behavior, then the clinician is almost certainly likely to do a better job than the statistician working with test scores. It is highly doubtful that a test could be made that could successfully identify the major habit patterns of the individual.

Tests are not well designed for predicting how a person will perform in particular situations of brief duration. Rather do they predict general characteristics of behavior over a period of time. It is generally much easier to predict how a student will achieve in courses over a four-year period than it is to predict how he will achieve in specific aspects of courses.

Problems of Multiple Prediction

So far in this chapter, consideration has been given to the problem of predicting a single criterion variable from one or more predictor variables. There are, however, more complex prediction situations

that must also be given consideration here. A common problem of multiple prediction is that of validating vocational guidance batteries for predicting vocational success. It clearly would not be practical to develop data for predicting success in each and every occupation, for it may be presumed that occupations can be grouped together into categories that call for similar combinations of abilities. The same may be true for predicting success in vocational training programs from this same battery.

The basic question in the problem that we have just considered is how many categories should be used in the classification of aptitudes for vocational skills. No very satisfactory answer can be found, because for two training programs to be classified in different categories it would mean that persons exposed to both would have to show a performance in one that was quite unrelated to performance in the other. Such a fact is almost impossible to establish at the present time, because it is not feasible to submit the same individuals to two extended training programs one after the other. Even if this were possible, difficulties would arise in maintaining motivation, and thus evidence would not be obtained of the relationship between achievement in these two learning situations. For this reason, certain indirect approaches to this problem have been proposed. One of these is to determine whether persons who successfully complete the two courses of training can be differentiated in terms of a battery of aptitude tests administered prior to training. If no such differentiation can be made, then the two programs are considered to belong to the same classification. This conclusion is based upon the assumption that all relevant aptitudes have been measured, which may not be the case at all.

The latter situation may be represented diagrammatically as shown in Figure V. In this figure, the aptitude scores of successful members of two occupational groups are shown with respect to two aptitudes, A and B. The members of one group are indicated by circles and the members of the other group by crosses. Neither aptitude alone provides a good discrimination between the two groups, but the two groups can be well discriminated by a function of the two aptitudes represented by the line YY′. This function, when it is the best possible one, is referred to as a *discriminant function*. A person's score with respect to this function can be used to indicate whether he is more likely to belong to the one group or the other.

When more than two groups are involved, there may be more than one way in which it is possible to discriminate one group from another. In such a case there will be more than one discriminant function. This is illustrated in a research by Tiedeman, Bryan, and Rulon (1953). In this study scores on seventeen tests were obtained for airmen in eight different Air Force jobs. The problem was to determine the extent to which the test battery as a whole permitted the discrimination of the men in the different occupational groups.

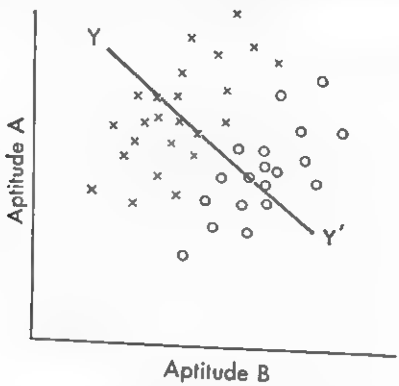

Figure V. Graphic illustration of discriminant function.

Two discriminant functions, each representing particular combinations of scores, showed some capacity for discriminating between the groups. One of these was a combination of scores that represented a variable differentiating mechanical and nonmechanical occupations. The other involved a differentiation in terms of the degree to which the occupation involved intellectual requirements. This illustration shows how more than one discriminant function may be found when many groups are involved.

It has often been considered that the discriminant-function procedure represents a desirable model for educational classification. When this is suggested, it is often forgotten that the mere fact that it is possible to discriminate between two groups does not mean that the basis for discrimination is necessarily one that can be used for future classification. An illustration may help in understanding this point. In a certain research organization with which the writer is

familiar, nearly all the research workers are men and nearly all the clerical workers and laboratory assistants are women. From these facts it is clear that sex may be used to make an almost perfect discrimination between the research workers and the assistants, but it is also clear that it would be most hazardous to classify future applicants for jobs as research workers if they are male and as clerical workers and assistants if female. Discriminant functions may be such that they are based on characteristics wholly irrelevant for purposes of subsequent classification. Many occupational groups may be discriminated from other occupational groups by accidents of history. The mere fact that a variable discriminates is not a sufficient basis for using it for subsequent classification. It is a necessary but not a sufficient condition for use in classification.

Most of the problems of prediction that have been considered up to this point are likely to resolve themselves ultimately into problems of selection. Thus we may seek to establish methods of predicting the classroom behavior of the teacher, but the ultimate justification of such studies stems from the fact that they contribute to an organized body of knowledge that will improve methods of teacher selection, and will enable us to discriminate between teachers who behave in accordance with some prescribed pattern that is considered desirable and those who behave in some manner that is inconsistent with this pattern. The usefulness of prediction studies in their contribution to selection and related guidance procedures has fully justified the extensive effort that has been channeled into them.

The Outcomes of Research on Prediction

The outcomes of research on the type of problems of prediction that have been considered in this chapter very largely represent contributions to the technology of education rather than contributions to an organized body of scientific knowledge. Nevertheless, when the outcomes of such studies are successful, they are likely to result in techniques for forecasting in situations where it is very important to make accurate predictions. Tests for the selection of college students, such as the American Council on Education Psychological Examination, were developed through prediction studies of the type discussed here. Although these studies, which were extended over nearly a decade, did not result in any major contribution to knowl-

edge, they did provide useful measures of certain aspects of scholastic aptitude. Nothing was discovered in these studies concerning how students learn.

Viewed from another light, and using Kenneth Spence's terminology, it may be said that in such studies the relationship between a response in a test situation R_1 and a response in a learning situation R_2 is discovered. Such relationships $R_1—R_2$ represent a low-level type of law that does not involve any direct functional relationship between R_1 and R_2. Whatever relationship exists is based on a complex ramification of events which are not understood at all by the discovery of the $R_1—R_2$ relationship. It is obscure relationships of this type that are unearthed and rendered useful by such prediction studies as have been considered in this chapter.

The Phenomenon of Shrinkage

It is necessary to bring to the reader's attention a phenomenon that is discussed extensively in books on statistics. This is the phenomenon of shrinkage, which is most easily understood through an example.

A researcher was interested in the personality traits that distinguished the most popular pupils from the least popular in the twelfth grade of a large high school. By means of a sociometric technique he was able to select two groups of one hundred pupils each. One group contained only pupils rated high by their peers in the matter of personal popularity, while the other included only pupils rated low. The researcher then administered a battery of sixty tests to both groups and computed the mean standard score of each group on each test. The five tests that gave the largest significant difference between the two groups were then considered to be those that could be used to discriminate between the two groups. However, on repeating the experiment in the following year, the differences in mean scores between similarly selected groups of pupils had *shrunk* to negligible proportions. This example illustrates the shrinkage phenomenon, but it needs to be interpreted in order to be understood.

What has happened is this: Suppose that the experiment were to be repeated one hundred times and that, on the average over the entire series, scores on test X differed negligibly for the two groups. Even if this were so, it is highly probable that in some of these

experiments there would be found substantial differences between the groups in their average scores on test X. This is the kind of thing that may have happened in the study involving the administration of sixty tests to the two groups of one hundred pupils.

In the illustration given, it is probable that the five tests selected as the most discriminating were those that happened to show a particularly large difference between groups on that particular sample. For this reason one can expect shrinkage of these differences on successive samples. The assessment of the true value of this difference could be made with some accuracy only by taking a very large sample of the two types of student. The fact that a large and unexpected difference occurs with a rather small sample is no basis for believing that it will turn up with another sample. The problem arises because the large number of variables involved permits the selection of those that happen to show particular "errors" in the study in question.

Just as differences in means may show shrinkage when an experiment is repeated, so too may other statistics show shrinkage when they are selected for extremes of magnitude and an experiment is repeated. Suppose that it were desired to predict success in shorthand courses in high school. A large number of tests might be given to beginning students in the hope that some of these might be correlated with later shorthand grades. If the tests that *happened* to have the highest correlation with shorthand grades were chosen as those most suitable for selecting shorthand students, the experimenter might expect to be disappointed when the tests were actually used for that purpose. Correlations of these "high predictors" with shorthand grades might be expected to shrink on subsequent samples. The reason for shrinkage in this case is exactly the same as in the previous example. Even if the tests when administered to a very large sample did not correlate with shorthand grades, on a small sample the correlations of these tests would be scattered over a wide range of values. The largest of these correlations would be large through the operation of chance circumstances that would be unlikely to be reproduced in subsequent samples.

A special and important case of shrinkage is manifested when multiple regression or multiple correlation techniques are used. Such techniques involve the determination of the best method of combining

two or more measures in order to predict a criterion. The best combination for a particular sample takes advantage of any peculiarities that make one set of weights more effective than another. Now when these same weights are applied to another sample that has different peculiarities, their effectiveness in predicting the criterion measure is reduced. Thus one commonly finds that when a multiple correlation coefficient is calculated on a particular sample, and the same regression weights are then applied to a new sample, the new multiple correlation has shrunk.

An example of the difficulties that may arise when the phenomenon of shrinkage is not taken into account is illustrated by a common technique of test construction. In the building of an aptitude test to predict a criterion, such as passing or failing a course in engineering drawing, it is practice to administer a great number of different test items to a group that is to complete the course, and then to select from this pool those items that are the best predictors of grades in the course. Under such conditions one may expect the predictive value of the items to shrink when they are used for predicting grades in future samples. If the number of items above a certain level of significance is equal to the number that might be expected on a chance basis, the selection procedure should be rejected as worthless.

Studies should be designed so that the hazards of shrinkage are minimized. If possible, the procedure should involve a method of estimating shrunken values. This can be done by a number of procedures:

1. Simple cross-validation procedure. This is the procedure that has been traditionally used. For example, in developing a test to predict a particular criterion, such as success in algebra, test items that seem likely to predict success are administered to a large group, say one thousand cases. The data are then partitioned into two sections with five hundred cases in each. Sometimes the partitioning is undertaken so as to make the two groups unequal, in which case the larger of the two groups is used in the first analysis. On the basis of the first sample of data the items most successful in the prediction of the criterion are selected. The hypothesis that the selected items will truly predict the criterion is then tested on the second sample of items. When the discriminating power of the selected items is determined on the second sample, the values will be found to have shrunk

from what they were when derived from the first sample, but the shrunken estimates of discriminating power will show considerable stability when applied to subsequent samples.

The problem of shrinkage is particularly acute when multiple correlation techniques are used. In most studies one should provide independent samples on which the original values can be checked.

2. Double and multiple cross-validation. In the design that has just been discussed, the values derived from sample A are checked on sample B. An alternative procedure is to derive two original sets of values, one from sample A and one from sample B. Those derived from sample A are then checked on sample B, while those from sample B are checked against sample A. In the case of an item-selection project it may be wise to choose for the final version of the test that is being built only those test items that stand up under both procedures. This double cross-validation method makes fuller use of the available data than does the simple cross-validation method.

3. Statistical methods of handling the shrinkage problem. For several decades statisticians have attempted to devise methods for estimating the amount of shrinkage that will occur on cross-validation. This problem has been most extensively studied in the case of multiple correlation techniques. The problem has not been solved in any satisfactory way, but it appears that certain iterative methods of computing multiple correlations offer some promise.

Nonlinear Relationships

Most prediction problems that are investigated within the domain of educational research are based upon the assumption that the relationships between the variables involved are linear. A *linear relationship* is simply one in which equal increases in the predictor variable are accompanied by equal increases in the variable to be predicted. It is generally quite acceptable to assume that any relationships that may exist are linear, for rarely have nonlinear relationships been found in the educational field, even when they have been actively sought. This is hardly surprising, since most measuring instruments are constructed in the first place to be such that they have a linear relationship with certain criterion variables. Thus the approach usually taken to instrument construction results in the lack of curvilinear relationships between the instrument and other variables. In

addition, those engaged in the study of individual differences have developed a wide range of statistical techniques based upon the assumption that relationships are linear. These techniques include those associated with factor analysis and such techniques as those of simple and multiple discriminant analysis, canonical correlation, multiple regression, and association procedures. By means of these tools the study of problems related to individual differences has been pursued. Such techniques are limited in the types of relationship that they can be used to study, and they are ordinarily quite unadapted to the study of nonlinear relationships. The adaptation of these methods to the study of nonlinear relationships usually results in the generation of functions that are extremely complex and require very elaborate arithmetical operations for their solution.

The problems of studying nonlinear relationships have been discussed to some extent in the section that pertains to profile analysis and pattern analysis. The latter techniques would be used when a relationship between two variables is not linear but the actual nature of the relationship is not known. The difficulties and impracticalities of utilizing most pattern-analysis techniques were noted, and it was also pointed out that the latter are likely to occur only when the researcher does not really know what he is expected to find. If he does know what to look for, then he can adopt techniques that look for this expected relationship and no other, and his task becomes a relatively simple one. Comprehensive pattern analysis looks for each and every type of relationship that might possibly exist; and since the possibilities are numerically large, the technique is correspondingly elaborate. The need is for studies that look for a few specific expected relationships, and then, perhaps, curvilinear relationships may be found.

Some Problems of Predicting Rather Rare Events

Meehl (1955) has pointed out that even though a measure may have predictive value for a given purpose, it may still happen that fewer errors may be made by *not* using it than by using it. Until it is understood, this paradox appears to present a situation filled with contradictions. Consider the problem of identifying persons who will become involved in delinquencies during a given year. Suppose that a test has been developed, which, it has been demonstrated, has value in identifying future delinquents. Let us also suppose that this

test was given to 10,000 high school children, and that 200 were identified as likely to become delinquent. At the end of several years, it would be possible to determine which of those identified as probably delinquent actually were delinquent. A table similar to Table 3 could then be drawn up.

TABLE 3. Hypothetical Data on the Identification of Those Expected to Be Involved in Delinquencies

	Number Actually Involved in Delinquency	*Number Not Involved in Delinquency*
Those predicted to be involved in delinquencies	30	170
Those predicted *not* to be involved in delinquencies	70	9730

One additional statement must be made in order to interpret these data, namely that the delinquency rate for this group is 10 per 1000. This is referred to as the base rate. With this fact in mind, the table indicates that the test does have some success in identifying those who become involved in delinquencies. However, by using the test on the group of 10,000, altogether 240 incorrect decisions were made (170 + 70). If no test had been given and if all of the group had been classified as nondelinquent, one would have expected only 100 incorrect decisions to have been made, which would have included all the cases that became delinquent. Thus fewer incorrect predictions are made by avoiding the use of the valid test than by using the test. Is it desirable to avoid the use of predictors where similar circumstances exist?

The answer to this is not a simple matter. Note that in Table 3 the test does identify correctly 30 of those later involved in delinquencies, but problems are created by the fact that it has erroneously identified as delinquent 170 cases who were not so. What has to be determined is whether the advantages gained by identifying the 30 delinquents outweigh the disadvantages of incorrectly identifying 170 as probably delinquent. If the testing requires an elaborate procedure and the help of many technicians, the losses may outweigh the gains. Also, financial and social problems may be introduced by identifying as potential delinquents those who are not.

The problem that has been discussed in this section becomes par-

ticularly acute as the base rate of the characteristic to be identified becomes very small. Attempts to identify rare talents, rare diseases, or any rare phenomena present situations such that selection devices are likely to provide a very much larger number of misclassifications than are provided by failure to use the instrument. This problem is most easily avoided when the base rate is near the 50 per cent mark. In addition, as the value of the test for selection purposes increases, the number of misclassifications is also reduced.

Summary

1. Studies designed to develop methods of prediction in education are of great practical significance but do not necessarily contribute to scientific knowledge.

2. There are three general approaches that may be taken to the problem of predicting educational achievement, which are as follows:

   a. The development of test situations that are miniatures of the learning situation in which it is desired to predict behavior.

   b. The administration of a wide range of instruments in the hope that one will predict.

   c. The development of a theory of prediction and the development of methods on the basis of that theory.

3. An event can be predicted if all of the conditions that ultimately lead up to that event can be observed or measured at the time when the prediction is made.

4. A major difficulty in predicting behavior is that one does not know the precise nature of the situation in which behavior is to be predicted.

5. The event to be predicted should represent a well-defined phenomenon.

6. Most prediction studies of behavior are based on the assumption that personality consists of a complex of relatively enduring and permanent traits. While stable intellectual traits can be measured, the same cannot be said of the field of personality.

7. A major difficulty in conducting many types of prediction studies stems from the fact that the conditions necessary for research on the prediction problem simply do not exist. This is particularly true of most attempts to predict teacher effectiveness. An example of a unique situation in which this problem could be studied with meaningful results was discussed at length.

8. Sometimes it is feasible to make a successful prediction in one section of a population but not in another.

9. A common method of improving predictions is to cluster variables that are related. Sometimes it is possible to cluster the criterion variables as well as the predictor variables.

10. Numerous studies have been made that attempt to contrast the relative success of the statistician and the clinician in making predictions from test scores. However, these studies do not compare the accuracy of the clinician in making predictions through clinical techniques with the accuracy of the statistician making similar or other predictions through psychometric techniques.

11. Many problems of prediction involve the determination of how many predictors are to be used for classifying persons into a number of groups such as occupational groups, and also how many of the occupational groups form a reasonable system for the classification of occupations. Research is only beginning to explore such problems.

12. In most prediction studies the phenomenon of shrinkage is likely to occur, and the studies should be designed so that it is possible to estimate the effect of shrinkage or to eliminate it.

13. One cannot always assume that the relationships between variables will necessarily be linear.

14. The fact that a variable has predictive value in a particular situation does not necessarily mean that it can be used profitably in that situation. Meehl's paradox occurs when the base rate for the occurrence of a particular event is low.

Some Problems for the Student

1. List some major assumptions that must be made in order to predict college enrollments fifteen years hence.

2. Identify some of the major difficulties involved in predicting the number of elementary school teachers that will be available for placement ten years hence.

3. One of the few skills taught in high school for which there are no satisfactory aptitude tests is typing. What hypotheses can be suggested to account for the failure to predict grades in typing classes from aptitude tests administered prior to training?

Studies of Development | 12

Studies of development, unlike most of those that have been considered up to this point, are concerned primarily with time trends, that is to say, with changes that occur as a function of time. Occasionally survey studies and prediction studies are concerned with time trends, but their techniques are not primarily directed toward the study of such phenomena. This chapter on developmental studies is included for two reasons. First, there are difficulties inherent in the conduct of such studies, which the student of education should know. Second, studies of development are of such far-reaching consequences for education that research of this type must assume a place of great importance in the future even if it has not in the past.

Some of the most important problems of education involve time trends, for education itself is concerned not with the static but with personal change and with the control of learning as it occurs in the pupil. While the classroom teacher is mainly concerned with change over a relatively short period of time, such as a semester or a school year, administrators and those concerned with policy-making at a high level may be concerned with change over a much longer period,

such as a decade or a lifetime. Classroom examinations given at the end of a semester are virtually limited studies in which increments in an area of intellectual skill are studied over a short period of time, perhaps without recognizing that change may not be permanent but will undergo a period of waning as well as waxing as time goes by. The latter type of problem has become of increasing importance as educators have extended their interest in educational problems from the childhood years to the entire life span. Studies of the intellectual functions in the later years are largely studies of decline.

Developmental studies may be quantitative or qualitative. Pioneer studies, such as those by Arnold Gesell on the development of motor and perceptual skills in young babies and those by Jean Piaget and his associates on the higher mental processes, have been essentially qualitative and descriptive in nature. Both have attempted to describe in words the nature of certain changes that occur in children as they grow. Little attempt has been made to measure these changes in these studies, but rather has the purpose been that of description. Perhaps it is necessary to begin developmental studies at the descriptive level, for then the investigator is able to obtain a feel for what is important to measure and what is trivial. Until the researcher knows what variables are of genuine importance, he is not really in a position to conduct quantitative studies.

While the value of descriptive studies is not questioned here, they might not be considered appropriate for a thesis or dissertation. Quantitative methods play such an important part in modern scientific methodology that the student may be expected to acquire some familiarity with them when he works on his thesis or dissertation. In addition, descriptive studies that do not involve measurement can so easily lead to false conclusions that they should be reserved for the experienced research worker. One is reminded in this connection of the astronomer Lowell, who thought he saw the canals on Mars as double lines rather than as single lines as earlier observers had described them. For many years thereafter astronomers confirmed Lowell's observation, but today it seems to be well established that the supposed doubleness of the lines was a figment of the imagination. Such are the hazards of qualitative observation. Perhaps it may be well to point out a few of the sources of these hazards.

First, the observation and description of behavior at various age

levels permits the observer to see what he wants to see—and even the best observers tend to some extent to see what they want to see.

Second, the observer faced with the innumerable events that constitute the flow of behavior is likely to be bewildered by the richness of the material. He may feel overwhelmed with the abundance of fact and feel that he does not have the kind of genius that can fit this bewildering array of events into a meaningful framework. As a result, he may direct his efforts to the matter of recording masses of material that he does not know how to handle.

It was pointed out in an earlier chapter that the qualitative studies of development over a considerable span of years, such as those undertaken by Gesell and Piaget, had their origins in the work of the nineteenth-century biologists who devoted many of their efforts to the study of life cycles and developmental patterns. Just as the biologist has tended to turn his interest in recent years to the study of conditions that affect development, so too have the educator and psychologist in recent years tended to turn their interest to the study of the relationship of environmental factors to development. Particularly has the latter become interested in the influences of the school environment on the pattern of responses acquired by the pupil. Such learning studies have obvious and immediate importance to the formulation of educational practices, but they have also attracted investigators because they can be undertaken in relatively short spans of time. These studies of short-term development related directly to the learning process will be considered first, since they are practical ventures for the student of education working toward a master's or doctor's degree. Later in the chapter, consideration will be given to the classical type of long-term study of development that finds its roots in the qualitative descriptions of development of the nineteenth-century biologists. A few such long-term studies do attempt to relate environmental conditions to learning. but the difficulties of doing this are substantial.

STUDIES OF DEVELOPMENT OVER SHORT PERIODS OF TIME

A great many studies have been undertaken that attempt to measure development over relatively short periods of time, perhaps as short

as a few days. These studies of development are particularly significant for education because of the emphasis that they place on learning phenomena. Such studies relate measures of development to environmental conditions and learning conditions on the one hand, and to the characteristics of the pupils on the other. They are designed to find out who learns what and under what conditions. In such studies, learning is conceived in the broadest sense. Not only are skills such as reading considered to be learned, but so too are thinking, judgment-making, and decision-making skills. Characteristics of personality are also presumed to be learned and subject to the influence of the school environment, and hence developmental studies of such characteristics as they emerge under different conditions constitute an important field for educational research workers.

Let us first give some consideration to studies of the development of intellectual skills, and follow this with a discussion of studies of personality development. Because of its brevity such a presentation can do little more than point up the merits and deficiencies of the methodologies involved. It cannot do what a textbook on development would attempt to do—provide a comprehensive review of the outcomes of research. The reader who wishes the latter can find many books that have been written for that purpose.

Short-Term Studies of Intellectual Development

Studies of short-term changes in behavior, if they are to be successful, are likely to be concerned with changes in intellectual skills. Illustrative of studies of this kind are the attempts to measure changes in thinking skills that were undertaken as a part of the Eight Year Study. In addition, one may point out that there are numerous studies of the acquisition of various aspects of reading skill, which form the basis of the present-day methods of teaching reading in the lower grades. Through such studies a vast amount of information has been gained about a great range of processes that enter into the development of mature skill. A similar range of studies, perhaps fewer in number, can be pointed out in the area of arithmetic skills. Most of these studies are not concerned with the mere mapping of the increments of these skills as they develop in our present-day culture, but rather are they concerned with the relationship of conditions of learning to the subsequent development of the skill. Studies that

attempt to discover such relationships are much more valuable in the information they supply than are studies that merely map the development of a skill under the varied conditions that happen to prevail at the present time. Studies that merely map a skill do not provide the information needed for improving instructional procedures in the classroom; they do not indicate to the teacher which of the various prevailing procedures are the best.

The reasons why short-term studies of the development of intellectual skills can be successfully undertaken, while those in the personality area cannot, need to be reviewed briefly, because they bring out some of the conditions that must exist for the successful completion of a developmental study.

First, the area of intellectual skills is one in which research workers have had over half a century of experience in developing measuring techniques, and as a matter of fact this half-century has produced a vast literature that provides an extensive theoretical basis for this technology. A great deal is known about the characteristics that should be built into measuring instruments if they are to manifest the desired properties. There is also a body of knowledge that enables the researcher to build these characteristics into actual tests.

Second, it is possible in the intellectual area to build instruments that have sufficient sensitivity to measure the small changes that occur in intellectual growth over the course of a semester. While such measurement may be quite inaccurate for individuals, it may nevertheless provide an accurate estimate of changes in groups, and most studies are likely to be concerned with group changes. In the measurement of intellectual skills there is little difficulty in increasing the sensitivity of a measuring instrument simply by increasing the length of the instrument.

Third, in the intellectual area, there is an extensive body of knowledge that has resulted in the development of theoretical systems of value both in the interpretation of findings and in the planning of further studies. Despite this fact, many studies that have been conducted concerning the development of intellectual skills have been carried out without any recognition that such a body of knowledge exists. To some extent this is excusable, because there has been a real need to collect empirical information concerning the relationship of broad and general learning conditions to the acquisition of in-

tellectual skills. For example, it has been of value to find out whether the excursion as an educational technique does or does not result in the increase of information possessed by the student, and whether it provides as much information and as well-remembered information as that resulting from other types of learning experience.

The trend in studies based on a theory of learning is for them to be based on some form of reinforcement theory though other types of theory are also used. *Reinforcement theory,* described in its simplest terms, is founded on the objective observation that certain events that follow a response increase the probability that that response will occur in the future. It is also observed that, when these events do not occur, the response is less likely to occur on subsequent occasions. The events that modify the probability that an event will occur are referred to as *reinforcers.* When the pupil reads the word "Boy" and the teacher says, "That's right," the words of the teacher are reinforcing events. Much that is done today in the study of classroom learning is concerned with problems of identifying and controlling reinforcing conditions so that desired pupil responses will occur with increasing frequency while undesirable responses will occur with decreasing frequency. A theory of this kind, which includes few additional concepts, may form a useful basis for studying such phenomena as the first stages of learning to read, but the more complex developments of reinforcement theory have to be introduced to account for more complex educational phenomena. Particularly difficult to handle are those aspects of learning in which the object is to learn not a single response, but a principle that is applied to the solution of countless different problems in the future. At this point the research worker is likely to find current learning theory of little value in structuring the known facts. Great care should be taken in the formulation of educational researches that study problems of development and learning in such complex areas. One may perhaps be reminded of some words of genuine wisdom given by Spence (1956) to remind the reader that priority in research should be given to those areas that "lend themselves to the degrees of control and analysis necessary for the formulation of abstract laws and comprehensive theories."

A major difficulty in the conduct of reinforcement studies in the classroom lies in the fact that educators have not as yet been able to

inventory or classify the reinforcements that operate there. For this reason, there is very little possibility at present for the experimenter or the teacher to exercise control over the reinforcing conditions. It is perhaps a reflection on the undeveloped state of the educational art that not enough is known to be able to tell the teacher how one of the most important conditions affecting learning can be controlled. The physician would be in the same predicament if he had almost no knowledge concerning the way in which diet affects the health and physiological functioning of the individual. The modern physician knows a great deal about the way in which diet can be manipulated in order to produce certain effects on the functioning of the body. The teacher lacks to a considerable extent analogous types of knowledge—which is the reason why it is common to refer to teaching as an art rather than a science. It may be assumed that, ultimately, such knowledge will become available. When this happens, it will no longer be necessary for the teacher to rely upon intuition to control the learning process, but rather will he be able to exercise deliberate and conscious control.

The central difficulty involved in the identification of reinforcing conditions stems from the lack of an adequate theory of the nature of human motivation. Most reinforcement learning theory has been developed by the study of animal learning in the laboratory. A considerable body of theory has been developed to describe such phenomena, but it does not seem to apply to human learning because it is so closely tied to physiological needs. Human motivation, at least after early infancy, is a much more complex phenomenon. McClelland (1953) and his associates have probably come nearest of any to the development of a theory that can be used in classroom studies. This theory would seem at this time to be an excellent point of departure for studying motivation and reinforcement in the classroom.

Much of the discussion up to this point has been concerned with the measurement of those skills that can be at least partially understood in terms of the better-known theories of learning. The acquisition of these skills represents only a limited aspect of the intellectual life of the child, though it is one that has been traditionally stressed. There is another area of intellectual activity on which school personnel have placed increasing stress in recent years and which is concerned with the processes of problem-solving, judgment-making, and

decision-making. This area of complex intellectual functions, together with the creative aspects of intellectual activity, have been particularly difficult for the educational psychologist to study, because until recently few theoretical constructs that provided productive approaches were available. During the last few years there has been a real change in this situation, and worth-while channels of research have been developed.

One important development has been the mapping of some of the abilities involved in complex intellectual activities. Much of this work has been undertaken by J.P. Guilford and his associates at the University of Southern California, who have had success in developing measures of the abilities involved. Another important development has been the extension of aspects of learning theory known as probability learning theory, and the application of these to the study of higher mental functions. A recent presentation of this approach, which has major research implications for education, is found in a work by Bruner, Goodnow, and Austin (1956). The argument presented by these writers is that man is able to cope with his environment because he is able to group together and categorize events, which then become discriminated from other events and other categories of events. Thus man has classified colors into a few categories, which are given names such as red, yellow, blue, etc. Thousands of colors that can be discriminated one from another can be classified under this simple system of names. A category is simply a class of events that are all treated as if they were equivalent. Much of human learning consists of acquiring the ability to discriminate what should be included and what should not be included in particular categories, such as dog, cat, and so forth. According to this theory, a person achieves *concept attainment* if he is able to discriminate between events that should belong and events that should not belong in a particular category. The studies by Bruner et al. are concerned with the conditions related to concept attainment.

In order to develop further the presentation of this type of theoretical development, consider the case of a child who is learning to discriminate between moving objects that adults categorize as dogs and those moving objects that are other than dogs. On seeing a moving object, the child makes a tentative prediction or *decision* whether or not the object is a dog. The decision is found to be correct or incorrect when he names the object and an adult indicates

approval or disapproval. This is the *validation* of the decision. The consequence of a decision is referred to as the *payoff*. The decision and the test of the decision provide potential information concerning the attributes that can be considered as predicable of belonging to the category dog. As this information accumulates, it progressively limits the possibilities of what is to be included in the category. The child might go about this by first attempting to make the discrimination in terms of color and by calling all moving *brown* objects dogs. This would be referred to as a *strategy*. In other words, the sequence of decisions through which information is acquired is called a strategy or a sequence of strategies. Strategies are retained or changed as they are successful or unsuccessful. The strategies adopted depend on the requirements of the problem situation.

This discussion is presented to bring to the attention of the reader the fact that a well-developed theoretical framework exists for conducting developmental studies of many aspects of the higher mental processes. Such studies conducted within a framework of theory will probably replace the purely descriptive studies that have been characteristic of the past.

Studies of Transfer and Generalization

The problem of generalization of learning, or transfer of training as it has been commonly called, is from the educational standpoint one of the most important areas of learning and development that can be investigated. If the school were conceived as a place in which isolated items of information were to be accumulated, the expected consequences of education would be extremely limited. Fortunately, this is not the case, for school learning is conceived largely as the learning of techniques, skills, principles, and methods that can be applied to a vast range of problems outside of the walls of the school. This is possible because the solutions to some problems learned in school are generalized to certain other problems outside of school.

Generalization is also invariably less than expected; that is to say, much knowledge tends to be compartmentalized. Some classic experiments have tended to show that some teaching methods are, perhaps, better than others in the extent to which the pupil is able to generalize the knowledge he acquires to other situations. This is an extremely important discovery, because the whole efficiency of education may depend upon the degree to which there is such generalization.

Despite the fact that there is some information concerning the effects of certain teaching practices on the amount of transfer of training shown, our ignorance is vast compared with our knowledge. A major difficulty in the design of studies related to conditions affecting generalization of training stems from the lack of any particularly useful theory, and until this deficiency has been remedied educational researchers are going to be hesitant about developing programs of research in the area.

Laboratory Versus Classroom Studies of Learning

The difficulties involved in the study of learning in the classroom situation have resulted in many research workers questioning the desirability of conducting experimentation in the classroom—at least not until much more knowledge has been acquired. There are two distinct points of view in this matter.

On the one hand, there are those who feel that meaningful experiments on learning can best be undertaken in the classroom, since the realistic conditions under which learning actually occurs can be found there. This view is supported by the persuasive argument that it is usually unwise and often impossible to generalize from the laboratory situation to the classroom. In addition, in the classroom children can be studied over a longer period of time than is possible in the laboratory.

On the other side of the question, it is pointed out that the laboratory has produced many important findings about learning. It is proposed that learning phenomena be explored first in the laboratory and, once positive results have been obtained, an attempt be made to reproduce the results in the classroom situation.

No definitive statement can be made concerning which approach is likely to yield the more information. The immense opportunities offered by each one of the two approaches suggest that both be explored. It is perhaps the research worker's disposition and interests more than anything else that should determine which approach should be taken at this time.

Short-Term Studies of Personality Development

Studies of the development of interests and attitudes have had a long history. These preliminary excursions into the field of personality development and the educational conditions that affect it were prob-

ably stimulated by the existence of instruments for the measurement of personality characteristics. L.L. Thurstone, who developed the first well-designed attitude scales, also initiated research on the role of various pupil experiences in the development of attitudes. Throughout the 1930's large numbers of studies relating attitude changes to educational experiences appeared in the literature. Such studies showed again and again that curricular materials designed to change attitudes did so in terms of responses to verbal attitude scales. There is as yet little evidence to show that the changes in attitudes measured by these scales are associated with corresponding changes in other phases of behavior. In recent years there has been some attempt to remedy this basic defect through the introduction of disguised attitude scales, which are designed in such a way that the person taking them is unaware of the purposes for which they are given. Scales of this type must still be considered to be experimental in character, and the relationship between behavior on such scales and behavior in other areas still needs to be established. There is also another question that must be raised about such studies; namely, whether the changes produced are not just changes in superficial characteristics that have little deep value for the child. Changing a child's expressed attitude toward a racial minority by showing him a film may mean only that, for the next little while, he will repeat the sentiments expressed there rather than those he has heard elsewhere.

The basic weakness in studies of attitude change as measured by verbal scales is that they are not based upon any particularly well-developed or well-accepted theory of personality. Modern conceptions of personality rarely even make reference to attitudes and interests; at least they are not considered major elements in the structure of personality.

While there is a wealth of theory on which to base research on personality development over relatively short periods of time, a major difficulty is presented by the fact that instruments for measuring such changes lack the sensitivity needed. This difficulty can be avoided to some extent by lengthening the period over which such studies are made, but this introduces difficulties to be considered in the next section of the chapter. Studies of the development of achievement motivation and the conditions that affect it have been carried out

successfully by this procedure. As far as the writer knows, there has been no successful attempt yet made to show how a stimulating teacher may produce increments in achievement motivation.

The guidance area is one in which educators have made a concentrated effort to produce personality changes of real consequence, and it is hardly surprising that this area has attracted many research workers interested in determining the changes that guidance programs produce. Early studies in the area were largely unsuccessful because of the lack of sensitivity of the measuring instruments available, and also because of the lack of a useful theory concerning the changes to be expected. When the present writer (1948) reviewed such studies, he could find no reported evidence that personality changes produced by guidance procedures were measurable. Recently, research methodologies have been developed that offer promise of measuring personality change resulting from guidance. Since 1948 considerable progress has been made in this area as a result of bringing together the theoretical position of Carl Rogers and the technical research skills of William Stephenson. The fortunate circumstances that brought these two men together evolved methods of studying the effects of counseling that have had great success in demonstrating personality changes resulting from the process. Stephenson, through the application of what he termed Q-methodology, was able to provide a technique for measuring important outcomes of counseling within the framework of the personality theory developed by Carl Rogers. The technique, unlike those previously used, was sufficiently sensitive to demonstrate changes occurring during counseling. This in itself constituted a major advance. Emphasis should be placed on the fact that successful developments in this area of research resulted from the bringing together of a theory and a technique through which the theory could be successfully explored.

STUDIES OF DEVELOPMENT OVER LONG PORTIONS OF THE LIFE SPAN

Studies of development over the life span have their origin in the work of the biologists of the last century. The study of life cycles had a prominent place in the curriculum developed by Thomas Henry Huxley for the training of biologists, and the influence of this cur-

riculum is still evident today. The study of the development pattern as it was pursued by nineteenth-century biologists involved the detailed description of changes in form and function as they occurred in the life cycle. This work had its origins in scientific curiosity, but, as often happens, important practical applications were soon found for the outcomes of this research. Information concerning the life cycle of small organisms became the means of controlling diseases such as malaria and yellow fever. Descriptive work on the development of organisms is still pursued today because of the important impact it has on problems of public health.

The work of the biologist in the field of development has been mentioned here because it has formed the foundation of the work that psychologists have pursued in this same area. The research of Jean Piaget is in the descriptive tradition of the biologist who has only very recently turned to experimental studies. Gesell's studies of the development of behavior in young children and Piaget's studies of the development of the higher mental processes are in the biologist's tradition of attempting to provide accurate records of changes as they occur in a living organism exposed to a particular environment. Figures VI and VII provide illustrations of work in progress in Piaget's laboratory. The modification of development through the modification of the environment represents a more advanced stage of educational inquiry. The fact that most long-term developmental studies in the educational field are of the descriptive type is no reflection on the research workers. It means only that they are still in the early stages of establishing a science. It is to their credit that they are the pioneers.

Studies of development over short periods of time such as six to twelve months can be undertaken by most research workers who have had a reasonable amount of training and experience, but serious practical and theoretical difficulties arise when longer periods of time are involved. This is particularly unfortunate, since all teachers and educational administrators must be concerned with the contribution that particular programs make to the lifetime development of the individual. The elementary school teacher should rightly be concerned with the achievement of immediate goals such as the acquisition of reading skills, but at the same time he cannot forget to consider the effect of his teaching on behavior as far remote as that of adult life.

One may hypothesize that the adult's attitude toward reading began its embryonic development in the elementary schoolroom, and that it showed a prolonged and continuous period of growth. Although it may be hoped that the educational process itself shows continuity, there is not much evidence to show that this is accompanied by a corresponding continuous process of development.

Figure VI. Child participating in a study of the origin of number concepts. Girl six years and two months old solving a problem involving the use of continuous quantities. The illustration is from a developmental study by J. Piaget and A. Szeminska (*La genese du nombre*; Neuchâtel et Paris, Delachaux at Niestlé, 1941). Photo by courtesy of Professor Bärbel Inhelder.

Long-term studies of development may follow either one of two rather distinct patterns. In one type of study, an attempt is made to follow a group as it grows and moves forward through life, and every attempt is made to retain contact with all members of the original group. This is the hard way of conducting developmental studies. In the second type of study, no attempt is made to follow a whole group, but individuals are selected for study at each one of several age levels. This may be referred to as the cross-sectional approach and

has the obvious advantage of permitting the completion of developmental studies without waiting for individuals to grow up. This second technique has had a long history of use. Its beginnings go back to the days of Francis Galton, who first made a systematic attempt to trace the growth and decline of human abilities. A brief review of

Figure VII. Child participating in a longitudinal study of the development of geometrical concepts. Boy six years old in the process of solving a spatial problem and manifesting somewhat typical difficulties. The illustration shows a situation from a developmental study by J. Piaget, B. Inhelder, and A. Szeminska (*La geometrie spontanée chez l'enfant*; Paris, Presses Universitaires de France, 1949). Photo by courtesy of Professor Bärbel Inhelder.

a classic study by him will be used here to reveal some of the weaknesses of the approach.

Galton's efforts to measure human characteristics and to trace their course of development not only represent pioneer attempts in scientific measurement but are also the earliest to cover the life span. The collection of data for these studies was made possible by the unusual circumstances presented by an international exhibition held at Earls Court, London, in the year 1884. Galton was invited to set

up a booth at this exhibition, and he seized upon the opportunity of using it as a means of collecting data about a sample of persons spread across the entire age span. With this end in view, he set up a number of tests, which covered such varied phenomena as height, speed of movement, the ability to make simple judgments such as those involved in bisecting or judging perpendicularity of a line, strength of grip, visual acuity, and so forth. The population on whom such measures were made consisted of whatever individuals happened to visit the booth—people of the sort who typically visits exhibitions, some in almost every age group, but a preponderance of those in the younger groups.

Galton tabulated the data in order to arrive at a general impression of the curve of each of the abilities measured, and for more than a generation his data remained unique in the field. A curve derived from such data is presented in Figure VIII.

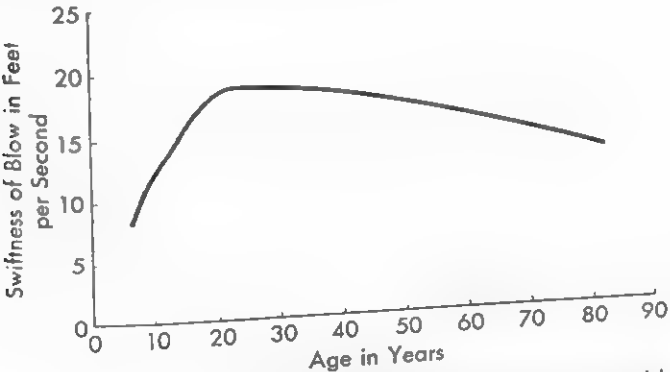

Figure VIII. Data illustrating changes with age. Relationship between age and swiftness of blow. Data collected by Francis Galton in 1884 and later reported by Ruger and Staessinger.

A straightforward interpretation of data such as those presented in the figure just considered is not as sound as it might appear on the surface. The curve cannot be considered to represent the typical growth and decline that an individual might be expected to follow. The reasons for this must now be given consideration.

First, the older group in a population is always a selected sample of what was once a much larger population. A group born in 1970

will lose some of its members through death by 1980, and death will not be an event that strikes at random. One can be quite sure that those who survive over the decade will be different in many respects from those who die. Among other things, the survivors will probably be living for the most part in more favorable economic circumstances than those who die. There will also probably be some selection for intelligence, in that the more intelligent will be more able to cope with the hazards of daily life. At the other end of the age scale there may well be selective survival for personality attributes that make for quiet and sober living. Thus in plotting the average of a given characteristic for each age group over the life span, one is plotting data for groups that are selectively changed at each age level. The data may be considerably different from those that would be obtained if the same group were followed through the years and if the measuring procedure were repeated at regular intervals. On this account alone, the cross-sectional method of obtaining developmental data must be considered highly unsatisfactory except for the crudest purposes.

An attempt can be made to exercise some control over the selective losses from the population as life progresses. One of these is to select groups at each age so that all groups are strictly comparable with respect to some important characteristics likely to be related to the phenomena studied. This procedure, known as that of obtaining structured samples, has considerable possibilities for the conduct of studies of development. Carson McGuire, for example, has an extensive program of research on development that involves this approach, and he believes it to be a most profitable one. Sometimes the attrition of time may be such that this kind of adjustment is not possible. There may be no acceptable way of matching a sixty-year-old group of women and a thirty-year-old group of women for education, since very few in the older group will have had a college education and it is not feasible to discard 90 per cent of the group.

Second, if the cross-sectional method is adopted in a particular community, the results may be distorted by selective migration. For example, the writer knows of a community on the northern border of the United States where in recent years there has been a tendency for the better-educated members of the young adult group to move south to the larger industrial centers. This migration has not affected

either those in the upper grades of school or those in more advanced years. If an intelligence test were to be given to every member of this community and the average score for each age group were plotted against age, the result would be a graph that showed a rise up to about the age of twenty, then a decline, followed by another rise in the region of the early thirties. A graph of this kind could not possibly be considered to represent the rise and decline of intelligence in this population, because other conditions are operating and determine the shape of the curve.

Third, even if the problems that have just been discussed exist only in minimal amounts, there is another difficulty in the interpretation of the data from the cross-sectional approach, which must be considered. In the case of the individual there is doubt whether it is reasonable to expect that development occurs in the manner depicted by the smooth graphs resulting from the type of data collected by Galton. Functions such as height may show periods of acceleration and depression, which result from disease or particularly unfavorable or favorable circumstances in the child's environment as well as from changes in glandular functions. These important individual irregularities in developmental functions are obscured by the cross-sectional approach.

The longitudinal method of studying development is not without its difficulties. However, these are not related to the interpretation of the data themselves, but to the process of collecting data.

First, there is the obvious difficulty involved in waiting year after year for the data to accumulate. This has not necessarily presented itself as insuperable, for many long-term studies that required researchers to follow cases over a decade or more have been made. Willard Olson (1949), for example, was able to follow the development of children on numerous variables as they progressed through school. He was not only able to plot individual curves for various functions, but was also able to interrelate the different functions and to demonstrate that changes in development rate are related to many differing conditions. In addition, Olson was able to demonstrate a certain parallelism between development in different functions, and he put forward the hypothesis that all development variables are functionally interrelated.

Second, there is the problem of attrition. It is not uncommon for

a researcher to find that less than one-third of his cases remain after he has been engaged in a developmental study for ten years. If funds for travel are available, it may be possible to follow those who have left the community, but this is an expensive process. In planning follow-up studies, the researcher would do well to choose a community in which a highly stable population, without much emigration, is known to exist. Whatever procedure is adopted, attrition adds immensely to the cost of the study, if only because this factor makes it necessary to start with far more cases than are to be included in the completed study.

Third, there is the difficulty of sustaining the interest and cooperation of the subjects in the study. This is not too much of a problem when the subjects are captive, as they are in experimental schools run by universities and colleges, but it is a problem when the subjects must be contacted at their homes and brought into the laboratory for study. Under such conditions, it may be desirable to reward the subjects with a knowledge of their performance and perhaps to provide financial rewards too, if the money is available. With adults, an effort may be made to keep the group informed of the purposes and progress of the study, but this attempt to stimulate interest will probably have appeal only to the brighter elements in the group.

The points raised here should indicate to the reader the time, money, and effort that are required for the successful completion of a longitudinal study of development.

Research on the Development of Individual Differences

In recent years a new type of study has appeared, designed to throw light on the development of individual differences. Textbooks on educational psychology published as far back as the first decade of the present century noted the fact that there is great uniformity of behavior at the beginning of human life. As the child grows, there is a constant increase in the amount of differentiated behavior that appears, and individuals show an ever increasing range of differences in the response that they manifest to the same situation. Thurstone was so much impressed with this fact that he proposed that a convenient zero of intelligence might be that point where individual differences vanished.

On most measures of intellectual ability, there is a tendency for individual differences to increase as age increases. Some caution

should be exercised in the interpretation of this fact, since the change in the spread of scores may be as much a product of the measuring instrument as it is of the function that is being measured. Instruments available for measuring intellectual skills may be such that they do not measure individual differences adequately at the lower levels. In other words, individual differences at all age levels may be identical. This interpretation is generally not considered acceptable, for it does not fit the common observation that young babies show little in the way of individual differences, their repertoire of behavior being limited to a small number of responses that all (or very nearly all) babies show.

A new approach has been taken to this problem in recent years through the extension of factor-analytic methods to this area. The reader is undoubtedly familiar with the fact that most batteries of aptitude tests that are administered to adults can be considered to measure a limited number of factors—the verbal factor, the numerical factor, the deductive reasoning factor, and so forth. It has been postulated by many workers that tests given in the lower grades do not manifest the existence of a series of underlying traits that these factors represent, but rather do they appear to demonstrate only the existence of a single general intelligence factor. The differentiation of behavior that occurs as education progresses is accompanied by the appearance of the distinct abilities that are measured by typical aptitude batteries. The procedure for studying this problem has been to administer batteries of similar tests at various age levels. The batteries used must be such that the tests differ in difficulty but not in content. Thus an arithmetic reasoning test would be one in which the problems at all levels involved reasoning with quantities, but differed from level to level in the complexity of the reasoning process involved. The computations involved should be of the same order of simplicity at all levels, since the computation aspect measures a different factor from the reasoning aspect. After such batteries have been administered to a series of different age groups, factor-analytic procedures are applied in order to determine whether the same underlying structure of abilities appears at the different age levels. While the evidence that is derived from such studies remains somewhat ambiguous, it does nevertheless tend in the direction of supporting the hypothesis that factors emerge in the structure of human ability as the individual grows up.

Looking Backward: A Technique for Developmental Studies

There are limited possibilities for conducting developmental studies by searching in the records for data that have already been collected on individuals. For example, many school systems administer intelligence tests at regular intervals through perhaps as much as the twelve grades. The researcher may look back through the records of the system and trace the development of individuals who have passed through it. With these data and a little patience, it may be possible to trace subsequent development and hence to have a record that covers two or more decades.

A major difficulty in following this procedure stems from the fact that school records rarely contain the material in which the researcher is interested. Also, school records in the matter of tests are far from being the reliable sources of information that one would like them to be. Too often, the records show the entry of an intelligence quotient without there being any indication of the test on which it is based. It is well known that intelligence quotients based on different tests are not comparable, and hence such data must be considered to be largely uninterpretable. The professional staff of schools should not be considered to blame for this situation, since they often are fully aware of its existence but unable to remedy it because of lack of the clerical help needed for the keeping of adequate records.

When data are available that adequately document the tests given as well as the scores, another problem arises. This is a result of the fact that different tests are often given at different times in the school program. If these data are to be used, it is necessary to convert the various test scores to a common base. The data for doing this may or may not be available. At the best, such converted scores are far from being as satisfactory as are scores based on comparable tests. The conversion process itself tends to weaken the data because it involves the use of constants that are only imperfectly estimated.

An unusually novel approach to problems of development has been evolved by D.R. Miller and G.E. Swanson (unpublished manuscript). These investigators were initially impressed with earlier studies that had demonstrated that some social and economic groups manifested a higher incidence of certain mental diseases than did other groups. For example, the lower socioeconomic groups produce more than

their fair share of schizophrenics. Now, while this is an important empirical finding, it does not indicate *why* this relationship exists. Miller and Swanson set themselves the task of investigating this and similar relationships for the purpose of determining the variables in the background of the individual that have important effects on certain aspects of later personality.

The general procedure adopted was that of identifying a sample of subjects to be studied, obtaining information about the child-rearing practices to which they had been exposed early in life, and then determining the response of each to a number of situations in which conflicts of motives had been aroused. In order to limit the number of variables operating, the selection of subjects was narrowed by confining the group to white Protestants who had been born north of the Mason-Dixon line. Additional restrictions included were that they had to be within one year of the average grade of boys of their age, emotionally stable, above average in intelligence, and from stock that did not come from northwest Europe. Those from broken homes were also eliminated. The application of all of these criteria resulted in retention for the study of only 1 per cent of the boys available. The extreme care exercised over the selection of subjects should emphasize to the reader the importance of this phase of a well-planned series of studies.

The mothers of all the boys included in the study were interviewed. In order to determine the way in which the boys responded to conflict situations, projective tests related to such conflicts were administered. These tests were all of the story-completion type and were designed to determine the way in which the individual responded to a situation involving a conflict of motives. The tests were also administered after the individuals had been exposed to a realistic situation in which the same conflict of motives was likely to occur.

The technique employed by these investigators is likely to be a productive one, and some of the relationships discovered are of considerable interest, even though they should not necessarily be accepted at their face value. For example, they found very definite relationships between the type of maternal discipline exercised and the way in which the subject expressed aggression. However, maternal discipline is not related to the severity of guilt feelings nor to the development of defenses against these feelings. Time of weaning was

found to be related to severity of guilt about death wishes (wishing someone were dead), about stealing, and about disobedience.

The major difficulty in the interpretation of the results of this series of studies stems from the possibility that the background conditions studied may not be the ones that are really producing long-term effects on behavior. It is possible, for example, that age of weaning may be related to a whole host of important attitudes on the part of the mother. These attitudes, in turn, may have important consequences on child development. Thus it may be that attitudes rather than time of weaning are producing the particular effects. The chief strength of the studies stems from the fact that the results fit rather well a widely held theory of the Freudian type.

Studies of Changes in the Later Years

Considerable interest has been manifested in recent years in research into changes in intellectual functions and personality characteristics in the later years of life. Much of this interest stems from the fact that during the 1930's the country was faced with the probability that the mean age of the population would show a steady increase as a result of the low birth rate. The implication of this fact was that there would be an increasing number of older workers, many of whom would need retraining for new jobs in a rapidly changing technology, and that there would be other problems of the utilization of the time of this section of the population. These problems of education, counseling, and recreation seemed to provide a new and intriguing field for research, although perhaps it must be said that they attracted the sentimentalists rather than the mature and established research workers. The poor quality of much that has been done in this area in the last fifteen years is probably accounted for by this fact. It has also been responsible for the endless array of questionnaires that have cluttered the field.

A central problem of considerable scientific and educational interest that has been studied in this area is that of the extent to which different abilities show decline during adulthood. There has been much speculation concerning the over-all changes in the functions of the nervous system over the entire life span, and Hebb (1949) has attempted to build a theory concerning these changes. Hebb's book

is a fascinating effort to coordinate and make sense out of a conglomeration of data, and it should be read by all who are concerned with changes in intellectual functions over the years. However, Hebb's theory does not provide a useful conception of any differential decline that is to be expected in the later years of life.

Empirical studies of the decline of abilities are extremely difficult to undertake for reasons that have already been discussed, such as differential mortality among different ability groups. There are also problems in testing older individuals, for such persons may be very poorly motivated in the testing situation and fail to show any enthusiasm for finding out their abilities. Insofar as this is so, the decline in measured ability may be at least partly a function of declining motivation. Just what abilities should be measured in the older adult in order to appraise his continued capacity to make a contribution to a community is also problematic. In the case of the young, it is common to study development in terms of those abilities that are known to be related to educability in the academic sense of the term, but one does not know whether these same abilities are related to the educability of the older adult. The problems of this area represent an important field of educational research, but one that should be approached with the greatest caution, for the difficulties involved in making genuine advances are immense.

A related problem is the determination of the age when man reaches his peak performance in particular activities. The importance of this problem to education becomes clear when it is pointed out that education should be planned so that the years of maximum ability are not those primarily occupied with course work. The latter may be the case at this time, for the tendency is for the doctoral degree to be completed in the late twenties, while in many areas the age of maximum performance may well be in the early twenties. The subject is not without problems from a research viewpoint. While the usual techniques have involved the determination of the average age at which the great men of history produced their finest works, one cannot deduce from the evidence that this age truly represents a peak of performance of the nervous system. Perhaps this peak of performance is mainly a product of the cultural conditions under which the persons involved happen to work.

Contrasting Empirical Data on Development with a Theory of Development

Most of the discussion of long-term development studies up to this point has been concerned with the problem of tracing the course of development as it occurs in contemporary society. This information is of only limited value, because it provides no indication whatsoever concerning the conditions that must exist in order for particular aspects of development to be maximized. Perhaps there is some value in knowing what is the typical pattern of growth, for then it is possible to know whether some observed behavioral phenomenon is unusual for a particular age group. The parent is then able to know much about the behavior to expect at each age level. However, despite this information, the parent is still at a loss to know how to handle deviations from the expected pattern, for data do no more than describe typical patterns.

It has been stressed in this volume that the merely descriptive represents only an initial step in scientific work, and it should be avoided except where it is necessary to start with the descriptive before more sophisticated levels of scientific knowledge can be acquired. A science on which educational procedures are to be based would have to provide information concerning the relationship of various environmental conditions to development. How this aspect of a science of education is to be built cannot be clearly seen from the present vantage point. Similar children cannot be subjected to different environmental conditions *over long periods* in order that the effect of these conditions on development can be studied. It is true that children can be found who are exposed to different conditions, but there can be no certainty that any differences are the product of those conditions, for the children may have inherited dissimilar characteristics before they were exposed to these conditions. It would be of immense importance to be able to separate the effects of these two types of determinants of development. This is a problem that has intrigued many psychologists and educators for many decades. Research in this area must now be considered.

Studies of the Relative Influence of Heredity and Environment

During the period from 1930 to 1940, numerous studies were undertaken for the purpose of determining the extent to which in-

telligence test scores could be considered to represent inborn capacity or the extent to which the environment had been favorable or unfavorable. One can well understand the circumstances that prompted such studies. The previous decade had resulted in the production of many widely used intelligence tests, and rather wild claims had been made concerning their value for estimating the intellectual capacity of the individual as it is determined by his genetic constitution. Some encouragement had been given to this view by the fact that Binet had painstakingly selected for his test those problems the solution of which did not appear to require much formal education. Although Binet himself made no claim that his test measured innate capacity, it was a short step for others to assume that the selection of items was such that the test measured some quite basic and stable ability to think. In a somewhat different category was the Army Alpha Test, which had a substantial loading with materials that are typically learned in school, such as those involving arithmetic reasoning and computation, but the interest of the 1930's was based much more on the type of test that Binet had developed than on the group-administered test.

Genetic studies of human intelligence centered around two main techniques. One of these involved attempts to change intelligence quotients upward by providing a highly favorable environment believed to be conducive to such change. The second approach involved a comparison of the test performances of identical twins who had been raised in different environments. Both of these approaches were soon found to present difficulties that prevented them from being considered as appropriate means of solving the problem that they had been evolved to solve.

A first criticism of these techniques is that they do not involve acceptable procedures of genetic research, which a systematic development of the subject requires. The professional geneticist is inclined to accept the view that the first step in the development of a science of human genetics must be the development of methods of labeling each of the chromosomes. A beginning has been made in this direction, and certain chromosomes have been identified, such as those that determine sex and those that determine certain blood characteristics such as the A, B, M, N factors. At this time geneticists do not have one identifiable characteristic carried by each and every chromosome, but they are aiming at that goal and are moving toward

it. Once it has been attained, a sound basis will have been laid for the systematic growth of genetic studies of human characteristics. However, that day has not been reached, and it is claimed that to plunge into genetic studies of highly complex aspects of behavior when the simple problems have not been solved leads only to ambiguous and perhaps uninterpretable results. Abilities such as are measured by intelligence tests are clearly extremely complicated phenomena, and, insofar as they are genetically determined, would require an extremely complicated genetic mechanism to account for them.

The general approach of the 1930's to the study of genetic factors as determinants of scores on tests of intelligence, which is based on attempts to determine the effect of training on such scores, seems almost to be based upon the incorrect assumption that what is inherited cannot be changed and that whatever can be shown to be modifiable cannot be inherited. Such an assumption is clearly quite unsound. One only has to be reminded that the defect of the foot known as clubfoot, which is clearly inherited, can be almost completely eliminated by surgery to see clearly that inherited characteristics can be changed by the environment. Even if a score on a particular test of intelligence taken at one particular point of development could be considered to be a measure of the intellectual capacity determined by the individual's genetic constitution, this does not imply that the limit cannot be changed by subsequent events. Certain experiences can affect intelligence test scores. This is clear when it is pointed out that direct training on the responses required by a particular test will inevitably raise scores on that test. Presumably related experiences will also produce some increment in the scores on a test. For example, playing with blocks may be expected to improve scores on test problems that involve the arrangement and fitting together of blocks. The real issue is whether the varied and differing learning opportunities to which children are exposed results in test score differences other than those that are attributable to genetic differences. The latter is particularly difficult to determine, owing to the fact that differences in genetic structure may also be associated with differences in the type of learning experiences to which the individual is exposed. Persons who inherit characteristics that favor the development of high-level intellectual achievement may have different home environments from those genetically less favored.

since they have parents who also have certain favorable genetic characteristics, and this in turn may result in differences in the learning experiences to which the individual is exposed. Human society results in the hopeless confounding of genetic and environmental conditions.

In addition, the concept that biologists have of the meaning of the term "environment" has changed. It is no longer considered to be simply a set of external conditions. There is also an internal environment, consisting of the fluids that bathe the cells of the body. There is no longer a clear-cut line of demarcation between the organism and the organism's environment, and, with the disappearance of this line, questions concerning the relative effects of heredity and environment have become less and less meaningful. Many today would take the view that such questions are pointless and that the issues involved are dead. They have been discussed here briefly because from time to time over the last fifteen years they have been revived, even though many consider them buried in the past. The educational research worker should at least be familiar with some of the reasons why it does not appear profitable to revive them.

The evaluation made here of studies that attempt to compare the relative effects of heredity and environment has been a negative one, but this should not leave the reader with the impression that the effects of external environment on development should not be studied. On the contrary, the writer is of the opinion that studies of the relationship of such conditions to development are of vast importance. For example, the studies reported by Hebb (1949), which are of this character, are of vast importance to education and have opened a new approach to problems of development. The thesis that emerges from the Hebb studies is that opportunities early in life to make certain discriminations permits the individual to acquire the ability to make these discriminations. The most dramatic illustrations of this thesis are provided by persons blind from birth who have gained vision in adulthood as a result of an operation. Such individuals have the greatest difficulty in learning the simplest visual task, such as that of discriminating between a circle and a square. While dramatic differences in early experiences appear to produce dramatic results that last a lifetime, one may hypothesize that less pronounced differences still have a marked effect. Studies of such differences in experience and their consequent effect on later behavior can be undertaken

without any reference to the old-time heredity-environment problem.

Summary

1. Since education itself is concerned with changes that occur over periods of time, studies of these changes are of great significance to the educator. Such studies involve difficulties of their own, which the educational research worker should know.

2. Studies of trends over short periods of time are likely to be learning studies, since other types of changes are apt to be too small to measure unless they are studied over a period of years.

3. Studies of changes in intellectual skills are those most likely to demonstrate positive changes over a short period of a few months, since techniques for developing such skills have been well worked out. Studies of changes in personality over similar periods of time are unlikely to be successful except in the area of attitudes.

4. Learning studies of central importance to education are those concerned with transfer. The main difficulty of developing work in this area is the absence of a useful theory of transfer.

5. Learning studies may be undertaken in the classroom or in the laboratory. Each one of these approaches has its own disadvantages and advantages.

6. Long-term studies may adopt two different approaches. They may proceed by following a group over a period of years, or they may include different groups at different stages of development.

7. The cross-sectional approach yields data that are difficult to interpret for a number of reasons. The older group is always a selected group, since death is selective, and so are immigration and emigration. In addition such studies cannot present individual differences in the pattern of growth.

8. Longitudinal studies also present difficulties in addition to the fact that they require long years of patient waiting. Attrition of the original group produces difficulties, and in addition there is the problem of sustaining the interest of the group.

9. Research on the development of individual differences through the application of factor-analytic procedures represents a new approach to the study of problems of development.

10. Developmental studies may also be conducted by searching for earlier records.

11. Studies of the decline of abilities are likely to provide results that are difficult to interpret.

12. Most developmental studies provide little information concerning the conditions that accelerate or retard development, except for those that are concerned with short-term changes resulting from classroom learning. Studies of the relative influences of heredity and environment have not provided much clear information that the educator can use on the limits of development. Such studies tend to be weak because they do not have behind them the genetic techniques needed to obtain meaningful results.

Some Problems for the Student

1. In recent years some businessmen have advocated a technique known as *brainstorming* as a means of obtaining new and original approaches to their problems. A group that brainstorms a problem first holds a session in which there is a free flow of ideas, and no idea is ever criticized during this session. The production of wild ideas is encouraged, and no critical evaluations are allowed. At a later session the ideas are evaluated and sifted for merit, with the purpose of isolating those that have practical value. The advocates of brainstorming claim that through this technique everybody can become creative.

Design a study in which an attempt is made to determine whether training in brainstorming techniques results in increased production of new ideas, and, if such an effect is found, whether it will transfer to areas of thinking other than that in which training has taken place. Assume that it will be possible to introduce training in brainstorming in high school classes.

2. A research worker was interested in determining whether curves of the development of different functions tended to be parallel; that is to say, whether an increase or decrease in the rate of increase of one function was accompanied by a corresponding increase or decrease in the rate of increase of other functions. He was able to administer tests of vocabulary and arithmetic reasoning to pupils who were starting the ninth grade, and administered parallel forms of the same test to the same pupils at the beginning of each subsequent grade in high school. The height and weight of the pupils were also measured when they returned to school each year.

What characteristics must the tests possess in order that the data can be used to test the hypotheses in which the scientist was interested? What units would be satisfactory for plotting curves for all functions for a single individual on the same sheet of paper? After the data, which had been plotted on graph paper, had been examined visually for the expected phenomena, what other methods could be used for examining the data?

Experimentation in Education 13

Terminology

There are certain terms used in research in the behavioral sciences with which the reader should be familiar. The person or other living organism whose behavior is studied in an investigation or experimental inquiry is referred to as the *subject*, or sometimes simply as S. In the literature of experimental psychology, S may stand equally well for a rat, a monkey, or a human being. It is an impersonal term denoting the living organism that is the center of the study. The person who conducts the investigation or manipulates the experimental conditions is the *experimenter*, a term commonly abbreviated as E.

The Meaning of Laboratory Experimentation

There is no clear line of demarcation between field studies and laboratory experimentation. We may, in a school that has a cooperative administration, divide third-grade pupils into two groups by a random procedure. We may then determine the relationship between spelling ability and varying amounts of daily drill (given, say, over

a two-month period). In such a study we are, in a sense, using the school as a laboratory, and in this laboratory we are *manipulating* a variable referred to as *quantity of drill*. When we speak of the *laboratory approach,* we usually mean that some variable of central importance to our study is manipulated. This is to be contrasted with the *field-study approach,* in which observations are made on events as they occur in situations over which the scientist exercises no control. The reader might well observe at this point that the spelling skill study just considered might have been conducted on a field-study basis. The scientist might have selected fifty classes approximately equal in spelling skill at the start of the third grade. He then might have obtained a measure of how much time was devoted to spelling drill in each one of the classes during a three-month period, and at the end of that time he might have measured the spelling skill of each pupil in each class once more. Thus an estimation could be made of the relationship between spelling skill and drill, but the knowledge thus achieved would be much less certain than that derived from the experiment previously discussed. It is possible that those teachers who provided the greatest amount of spelling drill were those working in districts where parents were most concerned about spelling and who were giving their children help in this skill in the home. Thus, in the field study, drill in spelling might be augmented by help in the home, and the observed gains might as easily be a product of that help as a product of drill conducted by the teacher. For reasons such as this the laboratory type of study, wherever it can be carried out, is likely to yield more useful and more certain information than the field study.

In the case of many studies it is desired to conduct in school situations, it is not possible to manipulate those variables it is wished to manipulate. When this occurs, if it is still desired to conduct the particular experiment, it becomes necessary to do so in the laboratory or in special buildings or facilities set aside for this purpose. When this is done, it is often necessary to simplify greatly the conditions that occur in the classroom, and the experimenter may decide to conduct his experiments with one pupil at a time rather than thirty.

There are those who raise the cry of artificiality when the proposal is made that studies be conducted under the grossly simplified conditions of the typical laboratory experiment. This criticism should

be evaluated in terms of the fact that most of our knowledge of the highly complex events of the physical world has been derived from the study of simplified events that the scientist has studied in the laboratory. The study of simplified phenomena under artificial conditions has been a highly successful technique in developing useful knowledge.

The term *experimental school* should receive comment at this time. Sometimes such schools are also called *laboratory schools,* but in actual fact they are not experimental in the sense in which the term is used here. Neither do they provide laboratory conditions, that is to say, conditions under which carefully controlled experiments can be conducted. The term "experimental" in this context refers more to the *novel* character of the curriculum and to the fact that something new is being tried out rather than to experimentation in the technical sense of the term. In this chapter, we are concerned with controlled experimentation in the laboratory sense rather than with the uncontrolled study of curricular innovations.

Experimentation may occur within the laboratory or outside of it. In the laboratory, studies are usually of a type that requires relatively small numbers of subjects and the careful control of many factors that cannot be controlled in other situations. When experimentation requires equipment or complex apparatus, it may be necessary to work within the laboratory. Of course the laboratory itself introduces variables, which it may be desirable to control but which cannot be controlled easily. For instance, human subjects who are introduced into a laboratory expect to behave in a certain way, or at least feel that the situation calls for certain kinds of responses.

Experimentation in classrooms is likely to be conducted when it is desired to use fairly large groups and where it is considered necessary to conduct studies of pupils in what might be termed their natural habitat. Of course, if it becomes known that an experiment is being conducted, this knowledge affects the behavior manifested by pupils and others. There are real difficulties in manipulating conditions outside of the laboratory, and some of these are considered in other parts of this chapter.

There is a great range of other situations in addition to the laboratory and the classroom in which experimentation may be undertaken. Indeed, any situation in which relevant conditions may be

manipulated can be considered as potentially one in which experimentation may be carried out.

The Need for a Cautious Approach to Experimentation

Experimentation is the most powerful method for deriving knowledge that has any certainty of validity, and hence it should be vigorously pursued. Nevertheless, experimentation is one of the more difficult of methods to pursue successfully. It is therefore necessary to consider in considerable detail all the common difficulties that experimental studies are likely to encounter. This is likely to give the student the impression that the difficulties of experimentation are so many and widespread that the new researcher should simply avoid experimental studies. Such an inference should not be drawn, but rather should the reader take the approach that once he is forewarned of the difficulties commonly encountered, he is well equipped to design productive experimental studies.

Before considering the difficulties involved in experimentation in education, brief consideration will be given to the role of experimentation in the development of any science. While it is commonly said that experimentation is the path by which a science advances, this should not be taken to mean that it is the only path of scientific advancement. Most of the major figures who have advanced science in the last hundred years have not been notable as experimentalists. Einstein never carried out a major experiment, and neither did Darwin or Freud. While experimental workers have checked many of the deductions of Einstein, these experiments followed rather than preceded major advances. To a considerable extent, experiments serve to consolidate advances already made rather than to be responsible for progress in and of themselves. Priestley's experimental studies of combustion served to demonstrate to the scientific world what he already was sure was true. Much classic experimentation serves the purpose of demonstrating to the world at large what the scientist already knows to be the case. The moral to be drawn is perhaps summarized by the statement that while thought without experimentation may be productive, experimentation without thought is futile. In other words, when the student embarks on an experiment, it is assumed that he is checking some aspect of a well-thought-out theory, which may be his own or somebody else's.

Let us now face squarely some of the major difficulties encountered in developing experimental studies, realizing that well-designed experiments can be carried out by the student who is aware of the common pitfalls.

Concerning Difficulties in Manipulating Certain Conditions

Some variables can be easily and successfully manipulated in experimentation with human subjects, while others cannot. In educational research, the most manipulable of useful variables are generally those related to learning conditions. Experimental research presents numerous studies in which comparisons have been made between groups that have learned by lecture alone and groups that have learned by lecture plus a moving picture type of learning experience. Other experiments have compared the usefulness of various cinematic techniques for producing various kinds of learning. Still other types of studies in which learning materials have been manipulated are those in which an attempt has been made to vary the characteristics of verbal materials by adjusting them to certain levels of difficulty as measured by one of the well-known reading difficulty formulas.

Manipulating training conditions. Another type of variable that is commonly manipulated in experimental studies is the amount of training of a particular sort to which subjects are exposed. An important type of experimentation that involves the manipulation of this kind of variable is that related to the problem of transfer of training, sometimes referred to as the problem of generalization. This problem, stated in its broadest form, is the extent to which one learning activity contributes to learning in other areas. Many important educational issues of the past and present are of this type. The old controversy concerning whether the study of Latin did or did not facilitate the learning of English was a controversy about a specific problem in the field of transfer of training. More recently there have been controversies about the effect of certain social studies materials on the social behavior of the children. In all such cases, the central problem is one of transfer of training.

In the experimental study of transfer of training, the variable that is manipulated is the amount of training. The usual design for such an experiment can be represented as follows:

Experimental Group	*Control Group*
1. Measurement of effectiveness in performing Task A.	1. Measurement of effectiveness in performing Task A.
2. Practice on Task B.	2. An unrelated intervening activity or rest.
3. Measurement of effectiveness in performing Task A.	3. Measurement of effectiveness in performing Task A.

In this design, the variable manipulated is Task B. In more complex designs, it may be possible to study the effect of varying amounts of Task B on Task A. The limitations of this design appear when it is necessary to introduce very large amounts of practice on Task B in order to have any hope of measurings its effect on Task A. For this reason, the age-old controversy about the usefulness of learning Latin has not been resolved by means of experimental studies designed to provide definitive and conclusive results. It has been necessary to study the Latin problem by indirect methods, which have been tedious and have brought results not always as clear-cut as one might expect to derive frome experimental studies. Lest any reader jump to conclusions, let it be said that there is no evidence for believing that the study of Latin is an efficient way to facilitate learning of English.

In the transfer type of experiment, the variable that is manipulated is not usually the type of subject-matter learning that occurs in schools. Usually much simpler phenomena have been chosen, and those that can be expected to have a transfer effect after only a limited amount of practice. Mainly for this reason, the experimental studies of transfer do not provide results that can be directly applied to school situations.

Environmental conditions that can be systematically varied within the framework of simple experimental designs are conditions such as lighting, temperature, number of pupils in the class, and similar straightforward matters. More complex conditions that can be manipulated are the presence or absence of visual aids of various kinds, characteristics of teacher behavior, and indeed characteristics of the learning procedures adopted.

In the simple design that has just been considered, no account was taken of individual differences, or *population variables,* as they are called. A more complicated design would have taken such variables into account. The experimenter or the person who designed the study might have wished to determine whether any transfer effort found occurred in dull pupils as well as the bright ones. The design could have been modified so that the data derived from it would have provided an answer to this question. More elaborate designs can be developed to take into account a number of population characteristics and to provide answers to questions concerning the relation of these characteristics to the variable that is being manipulated.

Manipulating motivation. Despite the success that psychologists and educators have had in experimental settings in manipulating variables believed to be related in some important way to the outcomes of education, there are other classes of variables of undoubtedly great significance that it has not been possible to manipulate successfully in experiments with humans. In particular, the manipulation of conditions related to motivation and stress has proved to be a task largely beyond the capabilities of present-day experimentation. The researcher in education should be familiar with the problems presented by this class of variable, and hence a brief discussion of this matter needs to be presented at this time.

In the case of motivation, much can be learned with regard to methodology from the way in which this is handled in experiments on the behavior of animals, although the type of theory of motivation based on physiological needs, which animal experimenters use, appears to have little application to human experimentation. In this area of research, it is customary to control motivation by means of deprivation; that is to say, an animal is deprived of food (or some other necessity) for a given number of hours prior to the experimental period. During experimentation, the animal's goal becomes that of obtaining the material of which it has been deprived. (This point does not have teleological implications.) It is important to note that motivation in all such work is measured by the amount of deprivation expressed in terms of time. Motivation is never measured in terms of how the animal actually behaves in the particular situation, for this would involve circular argument. One should not say that, because a rat is highly active in a maze, it is therefore well motivated,

and then use the concept of motivation to explain the rat's activity. As a general rule, it may be stated that we should never infer from the behavior to be explained variables that are then used to explain the same behavior. It is therefore necessary to measure the hunger drive in the case of the rat in terms of deprivation and not in terms of whether the rat appears to be hungry.

There are certain conditions, commonly described as *drives*, that have been used in certain psychological experiments but that so far have not been successfully manipulated in terms of deprivation. One of these is the so-called exploratory drive or curiosity drive. It is introduced as a concept to account for the behavior of an animal placed in a strange locality who shows substantial activity and behavior that in human terms would be described as exploratory behavior. However, at the present time it is not generally feasible to vary the amount of this drive or to manipulate it as an experimental variable.

In studies of human beings, not much purpose is served by depriving the subject of food and then using food as an incentive in subsequent experimentation. The search for food and the maintenance of a homeostatic balance are achieved by an indirect process in human society. The cues in the environment that arouse motivation are complex and are related in the most indirect way to the acquisition of the necessities of life, but it is necessary to manipulate these cues in the experimental situation. An example may perhaps clarify this point. In a certain experiment it was important to administer a test at a low level and at a high level of motivation. In order to do this, the experimenter told one group that the test was purely experimental and that nobody knew just what it measured. The other group was told that the test measured an extremely important ability that was essential for success in life. Experience has shown that on simple tasks variation in the orientation given to the subject produces differences in performance.

McClelland et al. (1953) assume that motivation can be aroused by the presentation of certain cues, as in telling a person that his performance on a particular task indicates his worth, but it is not assumed that all individuals respond to the same cues. The reader must be aware of the fact that some persons have so little regard for tests (or education) that their behavior is quite unaffected by any

implication made by the experimenter to the effect that a low score marks the person as a relatively worthless member of society. The result is that when cues are provided and manipulated in an experimental situation for the purpose of controlling motivation, only certain individuals respond to them. Since relatively little is known concerning the types of cues that are most likely to arouse motivation, it happens quite frequently that experiments are conducted in which the cues provided do not arouse motives in the subjects available (although they might arouse motives in other subjects). Until we know just what cues to use in particular situations and with particular groups, we are likely to conduct many experiments in which we fail to manipulate the experimental variables that we had hoped to manipulate.

A particular type of cue that has been used quite extensively in experimental work as a means of arousing motivation is threat. In a sense, in the experiment just described threat was introduced, since failure presented some threat, for to fail on the test was implied to indicate great personal inadequacy. In other types of experiments, threat may be of a physical nature, as when the subject is told that he is doing miserably, or that "most people obtain a better score on this test." Conditions that produce stress are to a considerable extent reproductions of real conditions that disturb composure, which occur frequently. There are many problems of great importance, which need to be studied, on the effects of such conditions on performance. Other related problems that need to be studied are those concerned with the training of persons to meet threat and stress with constructive effort rather than with confusion. Much needs to be done to reduce and even eliminate many of the threatening and stress-producing conditions that the school situation presents, although it is not realistic to believe that all such situations can be eliminated. Since some such situations will exist, it is desirable that efforts be made to train persons to meet them.

Manipulation, stress, and sources of anxiety. Although threat and stress, which is the response to it, are an important area for research, the results of studies have not been particularly profitable. The products of experimentation in one laboratory can rarely be reproduced in another laboratory. So inconsistent are the results that often when the same experiment is reproduced in the same laboratory, the data

may lead to opposite conclusions. The reasons for this were obviously not apparent when researchers first embarked on experimentation in this area, and not all of them have as yet been identified. However, there are some that seem to be sufficiently clearly recognized at this time to permit a brief discussion.

First, there is the difficulty presented by the fact that persons who visit a laboratory as subjects realize that no real harm will come to them, and that whatever threats they face will result in only transitory unpleasantness. On this account, whatever must be suffered by the subject may be accepted in much the same spirit as thrills and fears are accepted by those who visit the side shows at county fairs. The threats introduced by the laboratory situation may produce entirely different responses from those introduced by life situations.

Second, it is possible that most of the stress situations that it is desired to reproduce in the laboratory are not those that result from single incidents; rather are they those resulting from conditions existing over a relatively long period of time. There is at least a little evidence that neurotic conditions derived from childhood experiences are not the results of a single dramatic episode, but that they stem from recurrent situations that disrupt because of their frequent recurrence rather than because of their severity.

Third, the experimenter is limited for ethical reasons to the manipulation of certain mild threats. One cannot assume that responses to mild threat are the same as those to severe threat except in degree. It is quite conceivable, and in some cases we know it to be a fact, that the response to severe threat is quite different from the response to mild threat.

Fourth, in the laboratory situation the goals of the subject may be quite different from those in other situations where similar stresses operate. Unless some control can be exercised over these goals, the effect of stress on performance cannot be studied in any meaningful way, so the control of this aspect of the situation is crucial.

Fifth, experimenters are so limited in the stresses they can reasonably and ethically induce that serious questions may be raised as to whether it is worth even attempting to experiment in this area, despite its obvious educational importance.

Finally, stress, threat, and related variables can probably be studied most easily in the classroom situation as it ordinarily occurs. It is not

difficult to find classrooms where threat is frequently used and where the pupils live in a continuous state of tension. Such conditions probably represent much more severe threat and stress than the experimenter would ever dare introduce into any experimental situation.

Manipulating teacher behavior. Another type of difficulty arises in manipulating variables in educational research when it is sought to vary the behavior of teachers along certain dimensions. For example, a student may desire to study the effect of certain aspects of teacher behavior, such as the number of rewarding statements, on specific aspects of pupil learning. Teachers may be quite willing to cooperate and to provide a specified amount of praise for pupil accomplishment, but some teachers will be much more convincing than others when they praise a pupil. If an experiment has been set up involving a group of teachers who administer much praise and a group who administer little praise, the experimenter can be sure, however well he has trained and practiced the teachers in their respective roles, that some teachers will deviate markedly from the prescribed course of action. Whenever behavior is the condition to be manipulated, we cannot expect to conduct experiments with clear-cut results. Even more complicated and unsatisfactory are experiments in which the cooperating teachers are personally involved in the outcome and hence are likely to be influenced in their behavior by their own desires. Most experimental demonstrations of the merits of progressive methods in education suffer from this limitation, and particularly so because they usually require teachers who believe in so-called progressive methods to adopt, for experimental purposes, methods that they believe to be unsound. What happens under such conditions is that the teaching methods with which the progressive methods are to be compared are presented in a way that can be described only as a caricature. The results of such studies obviously cannot demonstrate any useful principle.

The main source of difficulty in controlling teacher behavior stems from the fact that, while the experimenter may wish to control the behavior, much of it is actually controlled by the teacher's own motives and his desire to see one outcome of the experiment rather than another.

However, there is another source of difficulty, which is the inability of certain teachers to assume certain roles. It is just no use asking some teachers to attempt to teach by methods that require them to

act as authority figures. These teachers simply cannot assume that kind of a role because it is inconsistent with their personalities and life goals, and because they do not have the repertoire of responses needed for playing the part. For similar reasons, other teachers are capable only of playing an authoritarian role in the classroom. Many teachers of the latter kind would feel threatened by the classroom situation if they could not maintain full control of it. Since this experimental difficulty is not always recognized by researchers, many experiments are undertaken in which the results depend more on the personal make-up of the participating teachers than on any other factor.

There is a final matter to be considered in classroom experimentation, and that is what may be called the personal bias of the pupils. To some extent, pupils will behave in the way in which they believe they are expected to behave. If they know that the class, or the teacher, is being observed, they are likely to cooperate with the teacher, since cooperative behavior is considered a most desirable form for children to manifest. This pupil phenomenon is a most pronounced one, and even teachers who have serious problems in maintaining class control may have no trouble when they are being observed.

Trial Runs as Explorations in Measurement

One of the major functions of a trial run is to determine what is and what is not measurable in terms of available instruments or new instruments that it is feasible to develop. Quite commonly an experiment or investigation is planned, but attempts to execute part of it demonstrate that the suggested procedure could not possibly yield any results because of the crudeness of measurement procedures. The need for such preliminary trial runs to establish the meaningfulness of results as well as the feasibility of obtaining measurements of adequate accuracy has not been properly recognized by educational reseachers. It would be easy to point to large educational investigations that have been pursued over many years at a cost of hundreds of thousands of dollars and that have produced no results of any consequence, yet these investigations would never have taken place if a few preliminary studies had been conducted. Such is true of most ambitious studies in teacher personality.

A common type of problem raised by a preliminary study is the

lack of individual differences in the field in which measurement is to be made. If all measures made have the same numerical value, then there is little point in the application of measurement, for measurement is a procedure designed to indicate how much phenomena differ from one another. An additional type of failure is due to the unreliability of a particular measuring instrument in the particular situation in which it is to be used. If it is a verbal instrument, it may be found to be incomprehensible to the particular group to which it is administered, and there may be little hope of modifying it while still retaining its original purpose. However, more frequent than any of these difficulties is the discovery in the preliminary trial that the phenomenon to be measured eludes measurement, even though attempts are made to adapt all available devices that seem appropriate.

Laboratory Analogs and Paradigms

Most experimental sciences advance knowledge about commonly observed phenomena by introducing into the laboratory simplified versions of these phenomena. Galileo wished to study the laws of falling bodies but found that their high speed under natural conditions made it almost impossible to study them, and also that natural bodies fell under such varied conditions that systematic study was difficult. For this reason he proposed to study bodies moving down an inclined plane. Such bodies move relatively slowly, and their laws of motion can be studied with relatively crude instruments. History has fully justified this practice, for the laws discovered through the use of such laboratory analogs or paradigms have been found to be a sound basis for making inferences about bodies falling under free conditions. When Count Rumford observed that the boring of cannon generated great heat, his inference that the heat generated was proportional to the work expended was one that could be tested only in a laboratory setting and with equipment other than cannon-boring machinery. Man's curiosity about lightning had to be satisfied almost entirely through the study of small quantities of electricity manifested by sparks in the laboratory. Cavendish, that prince of experimenters, could never have determined the density of the earth except through a laboratory technique that permitted him first to work out a value for the universal gravitational constant. The reduction of natural phenomena to laboratory-size paradigms, or analogs as they are

called, has been almost universally the main basis of scientific progress.

When Galileo decided to study falling bodies by means of bodies moving down inclined planes, he had a logical and rational argument underlying this procedure. He argued that the less the slope of the inclined plane, the less would be the force producing motion in the object. If the angle of incline of the plane was θ, then the gravitational force acting down the plane would be g sin θ, and the force acting at right angles to the plane would be g cos θ. On this basis, a simple mathematical function was provided to relate events in the inclined-plane situation to events in the free-falling-body situation. In contrast, in the behavioral sciences in general and in the educational branch of these sciences in particular, such well-established relationships between the laboratory phenomenon and the out-of-the-laboratory phenomenon do not exist. Such relationships as do exist can be expressed in words that are vague in comparison to the mathematical relationships characteristically found in Newtonian physics. Because the relationships thus expressed in words are vague, the generalizations derived from such laboratory studies lack the certainty of applicability to other phenomena that is characteristic of Newtonian types of generalization.

This means that the procedure for applying the laboratory generalizations of the behavioral sciences must involve much more caution than is necessary in the physical sciences. This does not mean that the physical scientist is never wrong in his field applications, for he is, but because of the rigorous nature of his rationale deductions he is less likely to be wrong than is the behavioral scientist. The rationale of the physical scientist can be wrong, and it often fails to take into account factors that influence large-scale phenomena but do not influence events in the test tube. For this reason, large-scale plants are sometimes failures although the small pilot plant was a success.

For these reasons, the results of laboratory experimentation with educational problems should be applied first to limited situations where careful appraisal can be made concerning the effects produced. Only when it can be shown that laboratory results can be at least partially reproduced in real-life settings should any widespread application be planned. At this point, great difficulties may often be

encountered, for the real-life situation may not permit the quantitative appraisals needed to give credence to the results of a field trial.

In spite of the risk that generalizations derived from laboratory experiments may not be applicable to real-life problems, many scientists feel that this should not deter us from experimentation with educational problems on a laboratory basis.

Apart from the obvious advantages of laboratory experimentation that have been discussed, there is the fact that many phenomena simply are not amenable to study under the conditions where they are ordinarily observed. This does not mean that all educational phenomena can be studied with advantage under laboratory conditions, because many are not amenable to such investigations. For example, if the researcher were interested in the effect of neurotic behavior of the teacher on pupil behavior, he would not use a laboratory approach, because psychologists would generally hold the opinion that the main effect of the teacher's neurotic behavior is observed after pupils have been subjected to it over substantial periods of time. In the laboratory, we could not and would not expose individuals to neurotic behavior over several months or years. Such matters must be studied in educational situations as they occur.

Some Difficulties in Undertaking Experiments

Problems of experimental design are now studied by mathematical statisticians as a comprehensive area of inquiry. The problems studied by this group are considerably different from those that need to be considered in this chapter, for they revolve largely around the efficiency of experimental design. This concept of efficiency is related to the matter of so planning experiments that the maximum amount of information is obtained from a given number of observations. Problems of experimental design in the latter sense of the term will be given brief consideration in the chapter that follows, which is planned to make the student of education sensitive to such problems and to perhaps encourage him to study further.

In addition to problems of efficiency, it is important that the student also be sensitive to certain difficulties in experimentation in the behavioral sciences that are largely a product of the type of events studied. These difficulties are rarely discussed in books on experimental design because such works are written mainly by statisticians

who are unfamiliar with common flaws in the mechanics of actual experimentation. It requires experimentation in the field to become aware of these difficulties, which are not necessarily a product of the logic of the design.

In the pages that follow, flaws that the writer has commonly observed in experimentation in education are discussed one by one. Undoubtedly there are many others that occur with lesser frequency.

Deficiencies in design due to failure to include a control group. This is the most elementary of all deficiencies in experimental design. A principal wished to find out how much progress his fourth graders made in social studies as a result of the curriculum offered. He was able to find a published test that seemed to measure the achievement of objectives of social studies stressed by the fourth-grade teachers, and he administered the test at both the beginning and end of the school year. He was pleased to find that the group made as much progress as that shown by the norm group described in the manual for the test. What the principal did not know was that pupils who did not study material related to the content of the test made just as much gain in score over the year as the pupils whose achievement was being evaluated. Experimental design always involves the establishment of conditions such that a comparison can be made between the effects of two or more conditions. Where the second condition is absent, the results become uninterpretable.

Deficiencies produced by the experimental procedure generating a variable. This deficiency is somewhat similar to that previously discussed. One should be on one's guard that the experimental procedure itself does not introduce increments in score that can be carelessly attributed to the experimental treatment. An example is necessary in order to illustrate this error in experimental design. Ballard (1913) performed a well-known experiment in which he assigned school children the task of learning poetry. At the end of the learning period, the children were asked to write out as much as they could remember of the poem. Next day Ballard returned to the school and asked the children to write out once more all they could remember of the poem. He was surprised to find that on the second occasion, the children were able to recall more of the poem than they were on the first occasion. This apparent increment in learning after formal learning had supposedly ceased became known as the phenomenon of

reminiscence, and for forty years it was described in textbooks on education and learning as a genuine phenomenon. However, information now available indicates that reminiscence is probably a product of faulty experimental design. The error lies in the fact that the procedure used to measure retention immediately after the learning session is itself a learning experience, which increases the scores achieved on subsequent measures of retention. Ammons and Irion (1954) performed an experiment in which groups were given poetry to learn. Some were tested according to Ballard's procedure, while others were tested only after an interval of time. Only those groups that were tested immediately after learning showed the apparent phenomenon of reminiscence. The groups tested after an interval of time produced average scores no greater than the average of the groups tested immediately after learning. This study suggests strongly that the supposed phenomenon of reminiscence is a product of faulty experimental design.

Various designs that can be used routinely have been suggested to take care of this type of hazard of experimentation. One that has been suggested makes use of four experimental and control groups and can be used generally for determining the effect of a particular learning experiment. It is especially suited to the type of study in which a pretest is administered to determine the state of learning at the beginning of the study, then a learning period, and finally a posttest to determine the state of learning at the end of the experiment. The four groups used in this design, denoted by the letters A, B, C, and D, are exposed to four different schedules as follows:

Group A —	Pretest,	learning experience,	posttest
Group B —	Pretest,		posttest
Group C —		learning experience,	posttest
Group D —			posttest

Only Group A is administered the entire series of tests and learning experience. The remaining groups are administered only varying portions of the schedule. In this way the experimenter can determine whether some irrelevant aspect of the experiment is producing any increment from pretest to posttest in Group A.

Deficiencies produced by contamination of data. Many experimental designs give spurious results because correlations are generated

by spurious elements. For example, a scientist was interested in discovering the abilities related to talent in a course in creative writing. As a part of his study, he administered a battery of tests of creativity to the students at the beginning of the semester and planned to study the relationship between these test scores and measures of the characteristics of their written products during the course. The tests were scored, and the researcher discussed these scores with the instructor in the course in order to obtain cues concerning the relationship of the tests to creative talent—but this was an entirely unfortunate mistake. What it did was to open the possibility that the instructor's evaluations of the students' writing might be influenced by his knowledge of the test scores. The data provided by the instructor concerning the students and their creative product was contaminated by the instructor's knowledge of their scores on the tests of creativity. In another example, a doctoral student of education was interested in comparing two methods of rating pupil performance but decided to perform both types of rating himself. In the latter case, both ratings were contaminated by the rater's personal opinions about the persons being rated.

Contamination is by far one of the commonest of the errors of educational research design that render data uninterpretable. Such contamination is often difficult to identify and may pass unnoticed. This is one of many reasons why research plans should receive independent review so that such factors can be identified.

Designs that make unwarranted assumptions about the nature of the scales used. The commonest examples in education of designs that manifest this error are those involving the use of growth scores. For example, a researcher set up the hypothesis that teachers who introduced into their classes rewarding comments (such as, "That's good, Billy") produced greater gains in pupil knowledge of social studies than those who did not. This study was to be conducted in sixth-grade classes in a large school system in which the teachers follow a rather rigidly prescribed social studies curriculum. The general plan of the study was to administer equated forms of a social studies test at the beginning and end of the sixth grade, and to correlate average gains in scores for each class with the observed frequency of rewarding comments occurring during visitation periods. If the researcher were not aware of the central defect of this design in the

early stages of his work, it would probably become apparent in the later stages, when it would become evident that some classes had greater knowledge at the beginning of the sixth grade than others had at the end. While some increased their average scores from, say, thirty to fifty items correct, others increased their average scores from fifty-five to seventy-five. These two increases are numerically equal and according to the design of the study should be treated as equal, but in actuality the equality of these two increments must be considered an unjustifiable assumption. As a matter of fact, there may be reasons for believing that the one increment is much more difficult to achieve than the other in terms of the time and effort required. Also, the two increments may differ qualitatively, in that the one may be achieved by bright students while the other is achieved by the dull. The two increments cannot be considered comparable, and studies assuming that they are should not be designed. Such studies will provide results that are uninterpretable.

Deficiencies that result when relevant variables are confounded with irrelevant variables. This is one of the more obvious errors of experimental design. Both the error and the meaning of the term "confounded" can perhaps be best explained by means of an illustration. A researcher wished to study the effectiveness of flash cards in the teaching of reading. In order to do this, sixty first-grade pupils were given a reading readiness test. They were divided into two matched groups such that for each pupil in one group there was a corresponding pupil in the other group who had the same reading readiness score and who was of the same age and sex. Both groups used the same readers and workbooks, but the teacher of one group devoted time to the use of flash cards each day while the other teacher did not. At the end of six months the reading skills of both groups was measured, and the relative achievements of the two groups were then compared. However, this comparison was quite meaningless, because any advantage attained by one group over the other might as easily have been a product of difference in teachers as a product of difference in method (flash cards versus no flash cards). It could be said of this situation that differences in teachers were *confounded* with differences in method, so that any differences in the two groups could not be attributed to the one or the other. It is imperative that such confounding of the main conditions should be avoided. This could have been done in the present case by extending the experiment to

other groups and other teachers. Duplications such as this are referred to as *replications*.

Deficiencies resulting from sampling by groups and not by individuals. Somewhat related to the previous error is this sampling problem. Consider the spurious design involved in a study whereby the effects of two methods of teaching reading were to be compared. In this study, the researcher selected from one school six second-grade classes that agreed to use Method A, and from another six second-grade classes that were to use Method B. The researcher drew the unjustified conclusion that the results showed that Method A was superior to Method B on the basis of the fact that, although both groups had closely similar initial scores on a reading test, the final scores for group A were substantially larger than those for group B. The conclusion was not justified because there might have been differences between the two schools other than those in teaching methods. Differences in socioeconomic level or social status of the two school populations might alone lead one to anticipate that differences in rate of learning to read would be found. What has happened in this experiment is that differences in treatment in which the researcher was interested have been confounded with other sources of differences. This is similar to the deficiencies previously discussed, but that it can be remedied without adding additional cases.

In the study that we have just criticized, the basic defect in the design could have been remedied by the simple procedure of dividing the six classes in each school into two groups, one of which would have been exposed to Method A and the other to Method B. In this improved design, it would be possible to estimate differences between methods within each of the schools and to estimate the differences between schools regardless of method. Assigning individuals at random to treatments rather than groups to treatments will always avoid this flaw.

Deficiencies resulting from failure of designs to take transfer of training into account. Most books on experimental design that are written in the Fisherian tradition fail to note a phenomenon, unique to the behavioral sciences, that complicates the problem enormously. This is the effect of transfer of training. Many educational experiments cannot be conducted by the efficient types of experimental design found in books on the subject because of this effect. An example of this difficulty is presented in a study by Thomas et al. (1956)

concerned with the problem of predicting trouble-shooting ability. Two examples of each of two types of mechanical problem were constructed, and it was hypothesized that performance on type A problems would be related to measures of rigidity because of the unexpected nature of the required solution, while problems of type B would not. The data were consistent with this hypothesis when type B problems were presented before type A problems, but when the reverse order (AB) was used, then performance on the first type B problem was also correlated with measures of rigidity. What appeared to happen was that when a type B problem was encountered after type A, the subject was set to look for an unusual solution. Under these conditions, a commonplace solution acquired the property of becoming unexpected.

Deficiencies due to insufficient cases. One of the most elementary errors in experimental design results from failure to include a sufficient number of cases, but no simple rule can be given to guide the student in this respect. Part of the difficulty stems from the fact that when very small differences between groups exist (in relation to their internal variation) more cases are needed to demonstrate the difference than when relatively large differences are involved. Much also depends on the nature of the experimental design that is used, for some designs are much more sensitive than others in identifying small differences.

However, quantity can never make up for quality in the collection of data. The researcher is always better off with a few carefully made observations than with large quantities of observations made under varying conditions and of doubtful reproducibility.

If very large numbers of observations have to be made in order to obtain a reasonably accurate estimate of a difference, then it is doubtful whether a difference of that particular magnitude is large enough or consequential enough for the researcher to spend his time in further studies of the phenomenon. The present writer's own prejudice, which he follows in his work, is that if a difference between two treatments is not clearly apparent when each treatment is applied to fifty cases, then the phenomenon is one of small consequence. Certainly phenomena for investigation can be found quite easily that provide more clear-cut results of the type sought.

Deficiencies in design due to failure to take subject bias into account. In most situations, there is a tendency for human subjects

to behave in a way that they feel is expected of them. Thus in a classic experiment in which a group was singled out for observation in a factory, it was found that any variation in the conditions of work produced an increase in output, which remained even after the original conditions were restored. Groups singled out for study in schools are likely to learn more than groups not thus identified. For this reason, in any educational experiment where there is an experimental group and a control, both groups should feel equally singled out, or better still, both groups should be unaware of the fact that they are participating in an experiment. For this reason, in experiments with drugs, one group receives the drug to be tested while the other receives a placebo made to appear and taste the same as the drug. Both groups are kept in ignorance of the fact that some received the drug and some did not.

Deficiencies introduced by the observer because he knows the treatment to which a subject or subjects have been exposed. In many studies in the behavioral sciences, the data are collected through human observers who must exercise judgment in the recording of their observations. These observers, much as they may try to act in accordance with the ideals of the scientist, have their own preferences concerning the way they would like to see an experiment come out. These preferences are likely to influence the data as it is recorded if this process involves any element of judgment. The best way of overcoming this source of error is to design the experiment in such a way that the observer does not know which human subject has been exposed to each treatment. This is not always possible, but in its absence serious reservations must be held concerning the way in which the results can be interpreted. This kind of difficulty is commonly encountered in the studies of progressive versus traditional types of classrooms, in which it is difficult, if not impossible, to hide from the observer the general nature of the school program. The observer has to look only at the paraphernalia available to know whether the school expects the teacher to run a "traditional" or a "progressive" type of classroom. This setting is likely to prejudice the observer into interpreting what he sees in a manner that will match the interpretation to the setting. This is a type of error to which even an observer who is aware of the problem is likely to be prone, and it is one that renders useless many attempts at systematic inquiry.

Deficiencies due to planning studies in which rare events form the crucial aspects of the data. An experimental design is not likely to be feasible if it is built around a rare type of event. An example from outside the field of education provides an illustration of a type of problem familiar to the reader. During the early days of the development of antipoliomyelitis serums, experiments were carried out in an attempt to determine the value of various experimental serums. In some of the first experiments, approximately 20,000 pupils were randomly assigned to two groups. One group was given the experimental serum while the other was administered a placebo. At the end of the season, when the incidence of polio in the general population had fallen to its lowest ebb, the number of cases of polio in the two groups were counted. In such an experiment, it might have been found that in the innoculated group there had occurred 6 cases and in the placebo group 10 cases. Now although this difference is numerically large, it can be accounted for in terms of the differences one might expect if many samples of 10,000 cases each had been administered the placebo. In the conduct of such research, it soon became quite obvious that what appeared to be large samples were inadequate for the purposes at hand; and it was necessary, as the reader will remember, ultimately to use samples of as many as 300,000 cases in both the experimental and the control group.

For example, suppose it were planned to introduce a safety program into the elementary schools of a small city. It might be proposed that steps be taken to evaluate the effectiveness of the program by excluding half the elementary schools from it and then by comparing the traffic-accident figures for these schools during a semester with those for the schools which had the safety program. The weakness of the design is that too few children are likely to be involved in traffic accidents for the comparison to be statistically meaningful.

Deficiencies in design resulting from the experimental procedure itself affecting the conditions to be observed. A serious difficulty in educational research results from the fact that the process to be observed is often changed beyond all recognition by the mere process of observation. The description and recording of events within the classroom presents this problem in an acute form. We can no longer accept the notion, based on wishful thinking, that the introduction of an observer into the classroom does not affect events therein, for

clearly it does. Indeed, some have suggested that it just may not be possible to study the events in the classroom under the conditions that ordinarily prevail. They have likened the situation to the Heisenberg principle in physics, which states that both the position and the velocity of certain particles cannot be determined at the same time. These difficulties of conducting classroom studies seem to be insuperable at the present time, but it must not be assumed that they do not exist. One should at least speculate on the effect that this difficulty may have on the result of studies.

The Availability of Appropriate Experimental Conditions

The preceding sections of this chapter emphasize the negative side of experimentation, the what not to do; but the mere avoidance of the worst pitfalls does not insure that the resulting experiment will be even mediocre in value. In the literature can be found study after study that are flawless in technique of design but otherwise completely inconsequential. Ingenious experimentation of the type that builds a science of behavior owes its contribution to the fact that it is built on a sound theory and that the idea could be developed experimentally under available circumstances. These two conditions need to be discussed further here.

A sound idea for experimentation in the behavioral sciences must find its roots in the current tide of organized ideas that constitute the present state of the art. Many ideas that appear sound from the viewpoint of the layman may not be sound from the point of view of current knowledge. The layman will always protest this statement, as he always has, for it is inevitable that he will conceive of himself as an authority on problems of education. This conflict between lay opinion and that of the scientist is not new and has occurred in fields other than the psychological. The layman's emphatic belief that the earth was flat or that it was the center of the universe are illustrations of common sense being wrong while the scientist was right. Today it is not uncommon for the student of education to base the ideas about which he wants to experiment on lay opinion as well as on his professional background. This is a real handicap, but it is hard indeed for a person who has spent the first twenty or thirty years of his life thinking in terms of the layman's conception of behavior to change and to think in terms of the scientist's conception. Early habits of thought are probably never entirely discarded.

An example of what is meant here may clarify matters for the reader. The author at one time observed some of his associates experimenting with a problem of teacher personality. The approach was that of introducing the prospective teacher into certain situations that showed systematic variation in such factors as the amount of stress, number of persons to be supervised, etc. The assumption was made that a person is what he is and will show the basic core of his personality in whatever situation confronts him. Another assumption was that the traits manifested in these experimental situations would be the same as those manifested in the classroom. This reflects a number of common conceptions (or perhaps one should say misconceptions) concerning the nature of personality. These conceptions are quite inconsistent with many phases of modern psychology, and particularly with those that recognize that behavior changes as the goal changes. Since the goals of the teacher in the classroom may be quite different from those of the same teacher in the experimental situation, it may be expected that behavior will correspondingly vary. Unfortunately, not enough is known at the present time to predict behavior in the teaching situation from behavior in the experimental situation, at least if one's point of departure is the layman's variety of trait theory. There appear to be wide individual differences in the ability to assimilate current technical theory into one's thinking and to utilize it as a basis for experimentation in education.

The second point made earlier in this section was that the effective experimentalist must have vision enough to see what can and what cannot be accomplished under possible experimental circumstances that present themselves. The difficulties of manipulating some conditions, such as those related to teacher behavior, have been discussed, as also have some of the problems of laboratory experimentation that limit greatly what can be done. The shrewd experimentalist will not disregard these difficulties and proceed as if they did not exist, but rather he will design experiments that circumvent them and that, as a result, permit the emergence of a clear-cut answer to the question that has been asked. This aspect of experimental design, almost as much as that of finding a problem appropriately oriented with respect to theory, calls for great ingenuity and makes the design of experiments a pursuit that calls for high intellectual power. There is no routine way of setting up experiments. Genuine contributions to

scientific knowledge are not made by the application of well-tried formulas to be found in textbooks, but rather do they require at least some small creative effort.

A Final Word of Encouragement

Finally, the student is again urged not to be overwhelmed by the difficulties of experimentation outlined in this chapter. Rather should he feel that he is now familiar with the major difficulties commonly encountered, and that he is now in a position to plan well-designed studies. Since most flaws in experiments arise simply because the novice in research is unaware of these types of flaw, the reader at this point should feel prepared to try his hand at designing experimental studies. The great value that this approach offers to the development of a science of behavior in educational situations is a factor that should urge him to use experimental methods whenever they are feasible. The more ambitious doctoral student may well deliberately choose these most powerful of all methods of collecting information.

Summary

1. Experimental studies differ from field studies in that the former involve the manipulation of a variable while the latter do not.

2. If experimental studies and field studies are undertaken under similar circumstances, it is likely that the experimental study will provide much more certain information than the field study.

3. Laboratory studies are experimental studies of a special type undertaken under conditions that are carefully controlled.

4. Experimental studies that involve the use of human beings are limited by the fact that they permit the manipulation of only certain variables.

5. Some of the most important variables in human behavior, such as those related to motivation, cannot usually be successfully manipulated in experiments with human beings.

6. When the condition that is manipulated is the behavior of the teacher, the experimenter is able to exercise only the most limited control over the variable that is to be manipulated.

7. Trial runs and exploratory studies are important steps in the development of experimental studies. They serve the purpose of determining whether the planned research is feasible in terms of the proposed techniques.

8. Laboratory studies are usually simplified versions of the phenom-

enon in which the scientist is interested. This approach can be justified in terms of the immense success it has achieved in the past. However, it may be much more difficult to generalize from laboratory studies in the behavioral sciences than it is in the case of the physical sciences.

9. There are certain common deficiencies in the design of experiments, which recur with such frequency that they should be familiar to all who undertake research in education. These are deficiencies resulting from:

a. The failure to include a control group when one is needed.
b. The experimental procedure itself generating a variable.
c. The contamination of the data.
d. The making of unwarranted assumptions about the nature of the scales used.
e. The confounding of irrelevant variables with relevant variables.
f. Sampling by groups and not by individuals.
g. The failure to take into account transfer of training.
h. The failure to include a sufficient number of observations to provide the precision needed.
i. The tendency of subjects to favor one outcome rather than another.
j. The human observer being biased in the making of his observation because he knows which subject or group has been exposed to which particular treatment.
k. The failure of the experimenter to recognize that he is dealing with a rare type of event.
l. The experimental procedure itself affecting the conditions to be observed.

10. Sound experiments are based on a sound scientific theory. The mere fact that a design is statistically sound does not mean that the experiment is sound.

Some Problems for the Student

1. An educator was interested in determining the relationship between intelligence test scores and grades in a certain school subject. In order to avoid contamination of the data, he himself administered an intelligence test and then kept the unscored answer sheets stored in a locked file until the grades were handed in. However, his data could still be considered as contaminated. Why was this?

2. One sixth-grade class was given no spelling drill during a semester, while another in the same school was given thirty minutes daily of spelling drill. Differences at the end of the semester were measured on a standard-

ized spelling test to determine the effects of the two procedures. List the sources of inadequacy of this design and then redesign the study.

3. A principal interested in safety education kept careful records over a number of years of the absences due to accidents among the pupils in a five hundred-pupil school. One year he decided to introduce a concentrated program of safety education, and at the end of the year he compared the record with those of previous years. Why would his data probably yield little of significance?

4. A research worker was conducting a study in which reading speed was to be measured under two different conditions. Under condition A the pupil knew the precise purpose of reading the material. Under condition B he was instructed to study the material because it contained information that would be of value to him later. The purpose of the study was to determine the effect of the specificity of goals on the amount of learning taking place. In order to carry out the study, all eighth-grade teachers were asked to release pupils for it over a two-month period. The pupils were sent to the experimenter one at a time. The teachers made the decision concerning which pupils should be sent. The experimenter first ran all subjects under condition A and then all subjects under condition B. What are the errors in this procedure?

Problems of Research Design | 14

The previous chapter presented a discussion of the practical problems of experimentation, with particular reference to the feasibility of undertaking various types of experimentation. The extended discussion reflected psychological as much as statistical design flaws in design. In this chapter the mechanics of design will be discussed, but throughout this discussion the reader should keep in mind that designs that are methodologically sound from the statistical viewpoint may be applied to trivial problems and may even be based on assumptions inconsistent with those that the use of the particular data require.

Terminology of Design

In order to understand research design methodology, it is necessary to understand certain terms that are commonly used in the discussion of designs. The knowledge derived from a research is generally derived from a *sample* of a *universe*. The *sample* might be all eighth graders in Chicago whose birthday fell on the first day of any month, and the *universe* might be all eighth graders in Chicago at the present time.

The researcher is sometimes interested in the effect of the presence or absence of some conditions on behavior, such as the effect of drill on spelling achievement, or the effect of knowledge of results or lack of knowledge of results on computational skill. Differences in the conditions in which the researcher is interested are referred to as differences in *treatment*. In the simplest type of educational study, differences between the presence or absence of a particular condition are studied, and this would represent a comparison between two levels of a particular treatment (presence or absence). In more complicated experiments, many different treatments may be involved and the interaction of these treatments may be studied. For example, one might study formal drill versus no formal drill in the teaching of mathematics, and the teaching might be undertaken by either extravert or introvert teachers. Extravert teachers might be more successful with drill methods than with nondrill methods, and the reverse might be true with introvert teachers. Designs that permit the estimation of the effect of each treatment can be adopted within a single study.

Sometimes the research worker is concerned with the characteristics of the individuals involved in a study, especially in relation to some aspect of performance. Thus in studies of the results of different teaching methods, the researcher may wish to take into account pupil differences in ability because he may believe that some methods are best with bright pupils and others best with the dull. The characteristics of the population studied that are taken into account in a design are referred to as the *population characteristics*. These may be physical characteristics or psychological characteristics such as are measured by tests. They may also be derived from the person's background and represent the type of experiences to which he has been exposed.

It is not the purpose of this book to familiarize the student with the statistical problems underlying advanced designs and their merits. Such matters are well taken care of in textbooks devoted to mathematical problems. However, it is believed that complex experimental designs are only rarely appropriate in education. They seem to have had their most extensive application in fields in which there is a large and well-developed body of systematic knowledge. Agricultural experimentation is typical of such a field. Much is known about the effects of various fertilizers, and there is a large and growing body of verified empirical information that provides essential background infor-

mation for advanced agricultural experimental design. Most applications of complicated research designs in education are related to matters of practical interest rather than to matters of scientific importance. Rarely are they seen in researches that attempt to contribute to organized scientific knowledge about behavior in educational situations. It is important to emphasize this point, since there seems to be a growing belief among educators that the best way to turn out graduate students who can undertake research is to give them a course in experimental design of the type based on recent advances in statistics. Such courses may teach the student about designs that are largely unsuited to research in the behavioral sciences. Such courses also fail to familiarize the student with some of the more important shortcomings of experimental design that are unique to the behavioral sciences. Too often the student of education leaves such courses intent on applying what he has learned but finding that the chief application is to routine studies, such as those that attempt to partition differences in pupil achievement among such factors as school size, rural-versus-urban differences, age groupings, sex groupings, ability groupings, and so forth.

There is some division of opinion among those engaged in educational research concerning the utility of complex designs that take into account a large number of different variables, except in areas where much knowledge has already been acquired. Those who design studies involving numerous variables claim that this is necessary if useful results are to be achieved. The argument is that many variables are involved in most behavioral phenomena, and hence these should be taken into account in any study that is planned. On the other side of the argument it is claimed that the research worker usually does not know what these variables are, and guesswork rather than sound theory is likely to be the basis for including those that are included. Only rarely do elaborate designs give the impression of being firmly rooted in theory. Skinner (1956), who has participated in this controversy, has pointed out that most of the important facts of science were discovered long before complex designs had ever been invented. In addition, it is true that important facts in the behavioral sciences continue to be brought out by workers in educational research using the simplest type of experimental designs. Many fine studies may illustrate the use of complex designs, but it seems likely that simple designs will serve a useful purpose for many years to come.

Functions of Statistical Methodology

It would be inappropriate in this book to provide any extended discussion of statistical methods, since these require extended study on the part of the student of education who is preparing himself to engage in educational research. The student will always be limited both by what he knows and by what is known in the field of statistics in the planning and execution of studies.

At this time it may be of value to review the major functions of statistical methodology as it currently impinges on educational research. Present methods serve the two broad purposes of testing hypotheses and imposing a structure on observations. The latter function is really an extension of that performed by descriptive statistics, which served primarily the purpose of describing groups of observations in terms of certain characteristics of that sample of observations. An example of the application of descriptive statistics is the summarization of the heights of a group of children in terms of the mean height of the group and the standard deviation of the heights. If the distribution of heights is known to approximate normality, then the mean and the standard deviation will specify and summarize the distribution of all of the scores. This function of summarizing large masses of data is still an important one, and it is used by all agencies that collect massive amounts of data for various purposes. In the last two decades, these methods have developed far beyond the simple methods and concepts such as were involved in the example just given and have entered a realm of great complexity. While the classical problems of descriptive statistics commonly involved the description of measures of a population with respect to a single variable, modern developments present multivariate problems, that is to say problems involving many variables. A battery of twenty tests may be administered to a population of high school seniors. It may be found possible to describe the data provided by these twenty scores in terms of a much more limited number of factor scores. Factor analysis serves this type of descriptive function and permits the summarization of numerous scores in terms of a few. In a sense it is possible to say that this procedure structures the data, and of course in a sense it does. By establishing certain hypothetical variables referred to as factors, it is possible to "understand" the remaining scores in terms of these factors. The factors do not neces-

sarily represent any underlying reality that has greater significance than the variables directly involved. Indeed, they may be strictly hypothetical entities introduced for convenience, and there may be no actual phenomenon that they may be said to represent with any directness. This should not be taken as an attempt to belittle the use of techniques that have been developed to give structure to data, for they have been of immense value in the development of the applied technology of testing and measurement.

Procedures that give structure to data have become progressively less and less acceptable as the core of doctoral dissertations in education and in the behavioral sciences. Part of the reason for this is that the procedures can be mechanically applied, but it is generally agreed that a thesis or dissertation should require the student to solve a problem at least partly by concentrated personal effort and reflection. There is little educational value in grinding out a study by mechanical means.

A second purpose of statistical methodology concerns the justifiability of inferences made from data, or, in other words, the confidence that can be placed in particular inferences. While it may be necessary to find a convenient structure for data in the early stages of an inquiry, as when ratings are combined to produce factor scores, the later stages involve the testing of hypotheses and the making of inferences from data. Tests of significance vary in the degree to which they are efficient. Some do not make proper use of all the information provided by the data, while others do. The validity of the tests, if they appear to be appropriate, depends upon the extent to which the assumptions on which they are based are satisfied by the situation in which they are applied. If a statistical test depends upon the assumption that the universe from which a sample is drawn is normally distributed, it may not necessarily provide correct information concerning the confidence that can be placed on a particular inference in connection with which it is used, if the universe is otherwise distributed. Usually we know only that the conditions required by most tests of significance are only partially satisfied, but some comfort can be found in the fact that, in the case of most tests of significance, empirical trials have shown that considerable departures from the assumptions called for can be made before the statistical test becomes substantially biased.

The graduate student of education is most likely to be mainly

involved in statistics that serve the second major purpose, namely the testing of hypotheses. Through such methods a science of behavior in educational situations is likely to be produced.

General Characteristics of a Well-Designed Experiment

First, it is necessary that the data of the experiment be free from bias. In testing the relative efficacy of two methods of teaching reading, it is not sufficient to choose two groups of schools that appear to be equal and to assign one method to one set of schools and the other method to the other set. Suppose these two sets of schools were the Southside Schools and the Westside Schools, and that the researcher decided to assign reading Method A to the former and reading Method B to the latter. In making this decision, he may have been influenced by certain quite unconscious biases. For example, he may have been convinced in his own mind that Method A was superior to Method B. He might also have given his prejudice weight in deciding to assign Method B to the Southside Schools, forgetting that previous test results showed that their pupils were generally slower learners than pupils in the Westside Schools. In such an event, the conditions of assigning teaching methods to schools were such that the data resulting from the experiment would inevitably show a bias in favor of Method A. It is just this kind of bias that a well-designed experiment is designed to avoid. This is accomplished by eliminating the influence of personal choice in the assignment of treatments to schools. If it were administratively necessary to treat each set of schools as a block, then the methods could be assigned to the schools by tossing a coin. However, to treat each set of schools as a block would be highly unsatisfactory. What is needed is to assign the methods to each school by use of a table of random numbers or by some other means free of personal bias. Some specific instances of introducing bias into data were discussed in the previous chapter.

Sometimes in the collection of data, bias is introduced by the fact that treatments are assigned on the basis of conditions over which the experimenter had no control. For example, the writer was at one time confronted with the problem of determining the effect of the use of diagnostic reading tests within a school system where some schools used these tests and others did not. Those that used the tests made some effort to interpret the profile of scores attained by each pupil and to plan a program of work designed to overcome the

deficiencies thus revealed. It was suggested by the school authorities that a simple method of studying the problem might be to measure the reading proficiency of pupils in the two sets of schools, and to determine whether the pupils in schools that used the diagnostic tests were superior in reading to the pupils who did not have the supposed advantages of the diagnostic tests and the related remedial training program. However, the results from such a study would almost certainly be biased. In the situation under consideration, it could be shown that the pupils in the schools using diagnostic tests came from more favorable home backgrounds than those in the other schools. This in itself would probably produce differences in level of reading in the two sets of schools and make the results of the proposed study uninterpretable.

This would probably not have been the only source of bias in that study. The schools in which the diagnostic tests had been introduced might have had better trained faculties than the other group of schools, and this superiority would have reflected itself in teaching and in the resulting level of reading skill. Perhaps this discussion may not only illustrate the problem of eliminating bias but also point up the advantages of an experimental method in which treatments are assigned to cases by a bias-free method.

Later in this chapter, consideration is given to a variety of errors that may introduce bias into an experiment. But the warning should be given here that bias has a way of creeping into experiments even when it is least expected. Some of the ways in which it can sneak up on the experimenter are discussed in later sections of this chapter.

Before leaving the topic of bias, some further explanation is needed. Students commonly ask the question, "Does the elimination of bias mean that if the effect of differences in treatment is zero, then differences between treatments will be zero?" The answer to this question is in the negative. If such is the case, then the observations based on the two treatments must be considered as samples from the same universe of observations. The means and characteristics of the two samples will vary from one another, as would be expected on the basis of sampling theory. Lack of bias means that they would not differ in any *systematic* way.

Second, the experiment must be designed in such a way that it is possible to determine the magnitude of the differences that might be due to sampling alone. This may be stated in another way; namely,

that the experimental data must yield an estimate of error. This condition could be overlooked in most of the experimentation conducted by the physicist or chemist, because in such experiments errors of measurement are extremely small and data tend to clearly support or clearly reject the hypothesis under consideration. It is only when the experimenter enters fields where errors begin to be large in comparison with differences between treatments that the concept of estimating error becomes a matter of prime importance. Many experiments of great significance were performed before the statisticians' concept of estimating error was introduced, but the scientists who undertook those experiments were not oblivious to the idea. They relied upon their knowledge of errors of measurement that their equipment and materials involved. They were able to make judgments concerning the significance of their results with a rashness that the modern behavioral scientist cannot generally afford. Nevertheless there are some excellent experiments in psychological literature where no systematic effort has been made to estimate error. The phenomena of stimulus generalization has been explored largely through studies in which there has been no attempt to estimate error, but here again, the results have been so clear-cut that the experimenter's knowledge tells him that errors are extremely small compared with experimental effects.

Third, the experimental design must insure that there is sufficient precision for the data to be able to provide answers to the questions that are asked. In an experiment known to the writer, students of education in their sophomore year were divided into two groups. One group was given extensive opportunity to visit school classrooms, while the other devoted an equivalent amount of time to additional academic work. It was hypothesized that those who had the school experiences early in their course would be able to profit more from the academic work and would be able to see its implications for classroom practice more clearly. The criterion for the success of this procedure was to be found in terms of the effectiveness of the students' performance in practice teaching. The experimenter was careful to divide the twenty-four sophomores by random assignment to the classroom visitation and the academic work groups, and fortunately all twenty-four stayed with the program long enough to complete student teaching. The students were rated during student teaching by the regular classroom teachers to whom they were assigned, on the

basis of their over-all effectiveness as well as on more specific aspects of performance. For the purposes of this study, the over-all rating of performance was used and an attempt was made to determine the significance of the difference in the rated performance of the two groups that had been exposed to different educational treatments. As one might expect, the results were negative, since the difference between treatments was small and the error term was extremely large. The experiment lacked the precision to answer the question that was asked.

What can be done to increase the precision of such an experiment? One answer, but perhaps not the most satisfactory, is to increase the number of observations. As observations are added to the original experment, a more and more precise and stable estimate can be made of the difference due to treatments. In the present case, additional observations could be added by dividing the sophomores year after year into two groups and providing the differential training.

But this is not the only method by which the precision of an experiment can be increased. It is the one that should be used as a last resort and only after other means of increasing precision have been exhausted. The main alternative involves the removal of that which can be ascribed to the effect of identifiable conditions from the error variance. Thus in the study of student teaching that has been described, the main condition affecting a person's performance may be his own childhood experiences in the classroom. The sophomores might be divided into a group who had been to relatively progressive schools and a group who had been to relatively conservative schools. An equal number of each of these could be assigned to each type of educational treatment during the sophomore year. In the analysis of the results, it would be possible to subtract from the error term that fraction of the variance resulting from differences in type of school attended. This would reduce the error term and thereby increase the precision of the experiment.

Traditional experimentation in the physical sciences involved the manipulation of a single factor at a time. Indeed, textbooks on experimental procedures that are more than twenty-five years old stress this aspect as an essential feature of the experimental method. Largely through the initial work of R.A. Fisher and the later work of his students and associates, the concept has been developed of varying more than one factor at a time. The advantages of such

multifactor experiments are numerous. First, they answer several questions within the framework of a single experiment. Second, each observation may contribute data to the answering of every question with almost as much precision as if the experiment as a whole had been designed for answering a single question. Third, through the multifactor experiment it is possible to answer questions concerning the effect of one factor on the other. This is a matter that was not easily investigated before Fisher developed his techniques, although the problem was a familiar one to many scientists. It thus may be possible to demonstrate that under certain conditions of work, incentives interact with the student's level of motivation, and thus it happens that the well-motivated student is the one who responds to the incentives for learning held out by the teacher. These interaction effects are probably extremely important in the behavioral sciences, although scientists are likely to remain preoccupied with more straightforward and less complex effects for the present.

Controls in Experimental Design

The design of research is closely associated with the use of what are called *experimental controls*. Although well-designed experiments and other forms of research have been undertaken for many hundreds of years, the use of the term by scientists goes back for only about one hundred years. Boring (1954), who has studied the history of the concept of control in experimentation, finds three common uses of the term, which have added confusion to writings on scientific methodology because they have been used interchangeably.

First, the term *control* is used to refer to a restraint on experimental conditions. Thus in the administration of a test to determine whether children who have had certain diseases suffer a hearing loss, it may be considered desirable to conduct the tests in a soundproof room in order that extraneous noises may not interfere with the results obtained by some pupils and not by others. Extraneous sounds are *controlled* so that the resulting conditions will be as uniform as possible.

Second, the experimenter exercises control over the variable that he is manipulating. In determining auditory acuity, sounds are presented that vary in loudness and pitch. It is important to control the pitch of the sound since some persons may have a hearing loss only for sounds of a certain pitch. It is known, for example, that as indi-

viduals grow older they begin to manifest a hearing loss for sounds of high pitch. Thus the experimenter *controls* the pitch as well as the loudness of the sound that is presented as a stimulus.

Third, there is a sense in which the scientist refers to *control* groups or *control* experiments. Boring introduces this meaning of the term by referring to Mill's methods of experimental inquiry. Mill's first method is the Method of Agreement, which states that if *A* is followed by *a*, then presumably *A* is the cause of *a*. The word "presumably" is used advisedly, since it is obvious that *A* is not necessarily the cause of *a* even if it always has preceded it. In my home, eggs are always served after grapefruit at the breakfast table, but nobody would claim that the grapefruit causes the eggs. In Mill's second method, it is postulated that if *A* is always followed by *a*, and if the absence of *A* is always followed by the absence of *a*, then *A* can be asserted to be the cause of *a*. This method is an extension of the first, and it involves the introduction of the control consisting of *the absence of A*. It represents a very common method of educational experimentation. For example, it can be shown that children who have certain speech defects improve if they are given remedial speech treatment and do not improve if such remedial work is withheld. If studies of this problem had demonstrated *only* that those who had remedial work improved, it would still leave open the possibility that improvement was due not to the treatment but to the passage of time and to various unidentified influences. However, by showing that the withholding of treatment is associated with an absence of improvement, the experiment is enormously strengthened and the conclusion that the treatment produces improvement may be justified.

Mill's method of *concomitant variation* is really only an extension of the two that have just been discussed. If this method were applied to the problem of determining the effectiveness of remedial speech training, it would be necessary to establish not two groups, one with and one without treatment, but a series of groups with varying amounts of treatment. It would then be possible to relate the amount of treatment to the amount of improvement in speech. There are certain obvious advantages of the method of concomitant variation over the simple methods. In the example under consideration, it would permit the determination of whether increasing remedial work beyond a certain point yields worth-while added increments in improvements.

It is quite possible that the pupil can benefit from only a limited amount of remedial work and that additional increments have little or no effect.

The reader may ask at this point why it is that the method of concomitant variation is only rarely used in *experimental* studies of behavior. Experimental studies in the literature are almost always designed to include only an experimental and a control group. Rarely are groups established with varying degrees of a particular type of treatment. The reason is purely a question of time and money. It is usual for the investigator to find difficulties in stretching his limited resources to permit groups of adequate size for both the experimental and control series. Intermediate series would involve prohibitive amounts of both time and labor.

The extent to which controls in the third meaning of the term need to be introduced is always a matter of judgment. If a teacher of calculus administers a pretest to his students to determine how much they know about the subject matter of his course, and then administers a final examination, and if the content of both tests relates only to the course in calculus, one may infer that substantial increases in scores from the pretest to the final examination may be attributed to learning in the course. Indeed, if this mathematics professor were to introduce a control group who took both examinations but who received no training in calculus, his colleagues would probably speak of him as being unreasonably overcautious and too free in wasting the time of his students. On the other hand, the psychology professor who also gives a pretest and a final examination may well wonder whether increases in scores are a result of learning in his course. In this case, the increase may be a result of general reading, discussions with other students, and related materials learned in courses in biology, sociology, and other subjects that overlap with psychology. In the judgment of the present author, it would be highly necessary to include a control group in the latter case in order to improve the possibility of attributing gains in score to the content of the course.

The student should also be aware of the possibility that a pretest may, in itself, be a learning situation. Although little knowledge of a field may be acquired through taking a test, the student may become familiar with the form of an examination and this in turn may facilitate the answering of the questions.

The Function of Replication in Relation to the Problem of Estimating Error

In the behavioral sciences, a single experiment in which a measurement is made on one subject exposed to experimental treatment and the same measurement is made on a control subject cannot yield meaningful results. The design of experiments that can achieve this end and that can be used to derive useful generalizations requires the introduction of what are known as replications.

The term *replication* is frequently used with reference to experimental designs, and it refers to the making of additional observations comparing two or more treatments. Some replication is obviously necessary if there is to be any experiment at all. This can be explained by an example.

Suppose that it is desired to determine the effect of a second-grade workbook on the development of skills. A very unsophisticated experimenter might start with two beginning second graders, of the same age and equal reading skill. One of these two pupils might be given the particular workbook to use during a semester, while the other did not have a workbook. At the end of the semester, both pupils would be again tested, and let us say the pupil who had had the workbook made the higher score. Just what can be concluded from such an experiment?

The answer to this question is that no conclusion of any value can be drawn. If two pupils have equal scores on a reading test at the beginning of a semester, they will probably have different scores by the end of the semester—as a matter of fact, they will very probably have different scores if they are retested only a day later. The latter effect illustrates the fact that there are errors of measurement. Thus the pupil who achieved the higher score might have achieved this score without the use of a workbook. Also the child who had the workbook might also have had certain advantages, such as a parent who worked with him on his reading difficulties. All of these uncontrolled sources of variation in final test scores are collectively referred to as *experimental errors*. In order to estimate their magnitude, it is necessary to replicate the treatments (with and without workbook) with additional cases.

This matter may be considered from another point of view. If the score of the pupil in the one group is X_1 and the score of the pupil

in the other group is X'_1, then the single difference $X_1 - X'_1$, cannot be evaluated for its significance because there is no standard with which it can be compared. If a second pair of cases is added to the data, a second comparison $X_2 - X'_2$ may be computed, but the added data also enable us to begin making an estimate of variability within each group through the comparisons of $X_1 - X_2$ and $X'_1 - X'_2$. As pairs of cases are added, it becomes more and more possible to evaluate differences between groups, because the data enable us to estimate what differences would be expected if both members of each pair were drawn from the same group.

The question is inevitably asked at this point concerning the number of replications that should be included in the design. This is not an easy matter to answer. Sometimes it is possible to compute the number of replications that are needed to attain a particular level of precision. There is also a second procedure, which is particularly applicable to research in the behavioral sciences, where data are collected not at one time but in a series of separate sessions. When this is done, replications can be added until the desired precision is reached; that is to say, until conclusions can be drawn with a definite degree of risk that they are wrong. A procedure known as *sequential analysis* may be used at any given stage in the collection of data to determine how many additional replications are needed in order to obtain the desired degree of precision.

Replication is necessary in order that the variability of subjects exposed to a particular treatment may be estimated. However, the multiplication of observations may serve an additional purpose if more is done than merely adding cases randomly selected in pairs from the same populations. In the case of the study of the two reading methods, it would be desirable to draw samples exposed to the two reading methods from different intellectual levels, and perhaps too from schools in different socioeconomic neighborhoods. If such a plan of investigation were pursued, it might then be possible to determine whether one method was superior to the other, not only for children in general but for children at different intellectual levels and for children from different socioeconomic backgrounds. An interaction might be found between method and intellectual level—that one method was better with the brighter children and one with the duller. If the design is properly planned so that it includes other factors, much more information can be derived from a single inquiry

than if all replications represent only the addition of randomly selected observations from the same population.

More information will be supplied by a single pair of observations if factors other than that which is being studied (differences in teaching method) are controlled. This statement should be qualified to the effect that control is important only insofar as it affects the variable in which we are interested; that is to say, in the example under consideration, reading achievement. The procedure may be adopted of matching one member of each pair with the opposite pair on one or more relevant variables. If pairs could be matched absolutely on all relevant variables, the remaining difference between pairs would be attributable entirely to differences in reading method. This situation is, of course, a limiting condition that can never be actually achieved. Usually we do not know what all the relevant variables for matching the pairs should be, and many that can be identified cannot be measured.

There is, of course, no certainty that even in an experiment where subjects are carefully paired for exposure to the two treatments there may still be differences between the groups thus selected for study. This is why it is necessary to obtain an independent measure of experimental error. All that can be done is to assign subjects to the two treatments in such a way that there is equal *probability* of the subjects in the two groups being affected by these uncontrolled conditions.

Matching procedures increase the precision of experimentation: that is to say, they increase the amount of information that can be derived from a particular number of cases. Thus experiments with matched cases can be undertaken with a smaller number of cases than when assignment is by a random procedure. Nevertheless, there are often serious difficulties in the matching of groups. Often relevant data are not available. Often there are not a sufficient number of cases to permit the careful matching of several groups. This is clearly shown by the fact that educational literature is remarkably devoid of studies that have involved the use of carefully matched groups.

There is also a criticism of matching procedures arising from the fact that matching has often been accomplished only after large numbers of cases have been collected in both the experimental and the control groups. When this is done, it usually results in the very uneconomic procedure of discarding cases, which means that time

and energy are lost. Such loss of data is not a necessary part of a matching technique, for if only carefully matched cases are used in an experiment that is tedious to undertake, it may be possible to reduce materially the time spent in experimentation.

Matching procedures have now been largely replaced by techniques that handle the problem by strictly statistical procedures. These procedures have also been an outcome of the work of R.A. Fisher, and they are known as *the analysis of covariance*. They are generally more efficient than matching procedures and do not make it necessary to discard cases, as often happens when matching procedures are used. While matching cases with respect to more than one variable is a difficult and cumbersome procedure, the analysis of covariance can be used to control differences on several variables, and thus it provides the experimenter with a powerful tool for exercising control over sources of error in his data.

Sources of Error

A basic problem in the design of research is the estimation of error. Without such an estimate, the results of a study cannot be interpreted. Little has been said about the sources of such errors, so a brief consideration of this matter is now in place.

A convenient classification of error is provided by Lindquist (1953). He divides sources of error into three types according to whether they are associated with *subjects*, *groups*, or *replications*. He refers to these three types of errors as S errors, G errors, and R errors, after the first letters in the words "subjects," "groups," and "replications." These errors can be illustrated by a simple example. Suppose that sixty first graders were to be used in an experiment to determine the relative effectiveness of two methods of teaching reading, and that they were divided at random into two equal groups. It is possible that one of these groups might have more than its share of bright pupils, and as a result that group would have an advantage in learning to read regardless of the method used. This source of error, which is due entirely to chance factors determining which pupils are to be exposed to each method, is referred to as an S-type error. However, even if the groups are perfectly matched, it is possible that one group might have advantages over the other group during the experiment itself. For example, it might happen that the one group had had a better teacher than the other. The errors introduced

through such uncontrolled events are referred to as G-type errors, since they are attributable to differences in conditions to which the two groups are exposed. If the same experiment were repeated in another school, it is possible that the method of teaching reading found most effective in the first school might be least effective in the second school, and this phenomenon might be a genuine one. It is certainly conceivable that a method of teaching reading that is highly effective for teaching children from literary homes might be a poor method for teaching children in impoverished neighborhoods. Such differences between replications are referred to by Lindquist as R-type errors. Such an effect as that discussed could also be referred to as an interaction of socioeconomic background and teaching method. It is desirable to design studies so that errors due to all of these sources can be taken into account insofar as they are relevant to the outcomes of the study. Here it is only possible to familiarize the reader with these various sources of error.

Factorial Designs

Up to this point in our discussion, consideration has been given to simple, classical designs that are based mainly on the concept that an experiment is performed by keeping all factors constant except one. The essential principle that R.A. Fisher introduced to revolutionize experimental design was the concept that more than one factor could be varied within the structure of a single experiment. Such an experiment involving many factors may contain all the information and provide the same precision as a series of independent experiments involving each of the facts singly, and it will provide savings in effort and work on the part of the experimenter. An additional advantage of the multifactor design is that it may permit the estimation of the effect of the interaction of the variables. The concept of interaction needs to be developed at this point.

The concept of interaction is a relatively advanced one in the history of the experimental sciences, and it has become particularly useful in the biological and social sciences. It has become a key concept in the biological sciences largely through the study of chemicals in the form of fertilizers and drugs. In the use of fertilizers, the interaction phenomenon is dramatic in effect. Nitrogen alone added to a deficient soil may produce little effect on growth, and the same may be true when phosphorus alone is added. However, when both

are added the effect on plant growth may be remarkable. Under these conditions, it would be said that the variance of plant growth due to nitrogen alone, or phosphorus alone, would be negligible, but variance due to the *interaction* effect of nitrogen and phosphorus would be large. Other well-known and important interaction effects are found in pharmacology, where the combined effects of drugs X and Y are found to be greater than what one would expect to find from the effects of the two drugs administered separately. This effect is known as the synergic effect, and it is extensively illustrated by the well-known procedure of compounding several drugs into a single dose of medicine.

In the behavioral sciences, it is not possible to point to clear-cut and well-recognized phenomena that illustrate the interaction effect, possibly because such interaction phenomena are not usually studied in most experiments. One reason for this is that it is only rarely possible to provide a rationale on the basis of which they may be studied. It is easy to see why plants do not grow in a deficient soil even if either nitrogen or phosphorus is added, for plants need both of these elements in an available form, and one can well understand why it is that both added together produce results greater than would be expected from the effect of each separately. Other interactions in the biological sciences are fully in accord with expectation, but in the behavioral sciences one can be much less certain of what to expect. Perhaps at this point it may be well to pause and consider some cases in educational research where one may expect to find interaction effects.

One such situation is presented by the relationship of teachers to the type of curriculum in terms of which they can most effectively work. It has been suspected for a long time that the teacher who presents what has been called an "authoritarian personality" has great difficulty in working within the framework of a school program in which pupil initiative is encouraged. It is alleged that such teachers work most effectively within a traditional type of curriculum, where nearly all activity is initiated, controlled, and directed by the teacher. Such a situation does at least call for habit systems that are consistent with those one might expect to find in the so-called authoritarian personality. On the other hand, the teacher who feels secure in a classroom situation, who does not find activity initiated by the pupils threatening, might be most effective in a situation

where there was no need to control every movement of the pupils and where he could function more as a counselor and guide than as a dictator. As far as the present writer is aware, it has never been demonstrated that there is this type of interaction between teacher and teaching program, yet it is reasonable to suppose that such an interaction may be crucial. If such a study could be undertaken, it should provide data crucial to the selection of teachers.

The above example is given just to indicate the interaction phenomenon in a meaningful context. Too often the interaction variances are measured and tested for significance without being given any particular meaning. Such a practice is not consistent with the development of a rational science of behavior in educational situations.

Interactions may be of any order; that is to say, they may involve any number of variables. It will be recalled that it is possible not only to study the interaction of variables A and B, but also to study the triple interaction of A, B, and C. However, the fact is that in educational research we usually have trouble enough in measuring the main effects without becoming involved in interaction phenomena. Many of the models that are commonly used in educational research specifically omit interaction variances and may even be based upon the assumption that such interactions simply do not exist. When tests to be used for guidance purposes are analyzed according to factor-analytic methods, it is assumed that their factorial component parts are combined in a way that precludes the interaction of these parts, and yet one can be sure that the interactions are extremely complex.

In the case of the comparison of the two methods of teaching reading, there were two methods which may be said to represent factor A in the study, two socioeconomic levels representing factor B, and two ability levels representing factor C. Information might be obtained from this experiment concerning each of the following:

The effect of A on reading achievement
“ “ “ B “ “ “
“ “ “ C “ “ “

The interaction effect of A and B on reading achievement
“ “ “ “ A “ C “ “ “
“ “ “ “ B “ C “ “ “

The interaction effect of A, B, and C on reading achievement

When reference is made to the interaction effect of A and B, it refers to the effect of combining those two factors over and above the effect of the two factors alone. Thus if there are three factors (A, B, and C) there would be a minimum of 2^3 observations that would have to be made in order that each one of these effects could be evaluated, since each level of A would have to be combined with each level of B and each level of C. In actual fact, there would have to be a replication of the eight observations in order to increase the precision of the experiment to the point where it could yield useful information.

Where the number of factors in an experiment is 3, the number of observations needed, without replication, is 2^3. With n factors, it is 2^n. It can be seen that this becomes numerically large very rapidly as n is increased. With 10 factors, it would require 2^{10} observations, which is equal to 1024. This assumes that the effect of all of the interactions is to be computed. Replication would also be necessary in most studies in the behavioral sciences, and thus the number of observations to be made would be extremely large and might even be larger than the number that could be made with the time and facilities available. In such a situation it is not necessary to go back to the old system of varying one condition at a time, for the newer principles of experimental design offer certain compromises that reduce the number of experimental observations needed. In this latter type of design, application is made of a procedure known as *partial confounding*. In this procedure, a design is set up so that it is possible to estimate the effect of only some of the interactions. However, since most of the interaction effects cannot be identified with any known or hypothesized phenomenon, their effects are pooled in order to provide an estimate of error. The result of partial confounding is to cut down on the number of interactions from which an estimate of error can be made. The concept of partial confounding is extremely valuable in the development of efficient experimental design.

An important example of a partially confounded design is the Latin square, first introduced into agricultural experiments, where its meaning can be easily understood. Consider the case of the agricultural experimentalist who wished to compare four different fertilizers that varied only in the amount of phosphorus they contained. Let these

fertilizers be labeled A, B, C, and D. This experimentalist had available a square plot of land, which he had divided into sixteen equal smaller plots, each one of which was also square. His next problem was that of deciding which treatment to apply to each plot. The plots are shown in Figure IXa.

Any assignment of treatments to plots must take into account the possibility that the soil may show greater fertility on one side of the total plot than on the other, and the Latin square takes into account just this possibility. The assignment plan shown in Figure IXb represents the application of one of many possible Latin square designs that might be used to solve this problem. It should be noted that each treatment occurs only once in each row and only once in each column. There are many other Latin squares that satisfy this condition, and the one utilized should be selected at random.

a. Plot arrangement

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

b. Assignment of treatments

a	b	c	d
c	d	a	b
d	a	b	c
b	c	d	a

Figure IX. Illustration of Latin square design.

An example of the application of a Latin square problem to education may now be considered. Let us reconsider the problem of determining the relative effectiveness of four methods of teaching reading, denoted by the symbols A, B, C, and D. It might be possible to set up four methods of instruction that differed in the techniques used. The research worker hypothesized that children in some school districts learned to read more rapidly than children in others, and that perhaps the workbooks used also made a difference. Thus it was planned to conduct the experiment in sixteen schools which varied both in school district and in workbook used according to the scheme shown in Figure X. Immediately below is Figure XI, which shows a method of assigning the treatments to schools. The assignment is

such that all four treatments appear in each school district as well as with each workbook. It is then possible, once the data are collected, to compare school districts, to compare schools that differ in workbook, and also of course to compare treatments.

It would be quite unthinkable to select a single case from each one of the sixteen schools, since our general knowledge of educational measurement would tell us that such an experiment would be so lacking in precision that little if any useful information could be derived from it. The researcher on this account would probably select as many as thirty pupils from each school, which would amount to having thirty replications of the Latin square design.

		SCHOOL DISTRICTS			
		1	2	3	4
WORKBOOK	Type 1	School 1	School 2	School 3	School 4
	Type 2	School 5	School 6	School 7	School 8
	Type 3	School 9	School 10	School 11	School 12
	Type 4	School 13	School 14	School 15	School 16

Figure X. Diagram showing the selection of schools for experiment in terms of school district and workbook used.

A	B	C	D
C	D	A	B
D	A	B	C
B	C	D	A

Figure XI. Assignment of treatments to schools according to Latin square design.

It should be noted that this design answers a very general question as to whether the methods can be considered to produce different results. It is not designed specifically to examine the question of whether one particular method is, as suspected in advance, superior to the other methods, though such a hypothesis can be tested. The design is perhaps appropriate if a very general exploration is to be made of reading methods, merely to see if different methods do produce different results.

In recent years considerable interest has been shown in the prob-

lem of testing the significance of particular comparisons within the Latin square once it has been demonstrated that there are over-all differences of significance among the treatments. Some of the methods for doing this involve quite elaborate assumptions, which should be fully recognized before these methods are embarked upon.

If a first principle of design is that there must be a way of estimating the probability that an observed difference between treatments could have resulted by sampling from a single universe of observations, a second principle is that designs must be arranged so that known sources of variability can be separated from both the main treatment and the estimate of error. Thus Mitzel and Rabinowitz (1953), in an experiment on the social-emotional climate in the classroom, employed a design that permitted the estimation of differences between observers, differences from day to day in the performance of the same teacher, and differences between teachers, which was their center of interest. Consequently they were able to remove from their estimate of error variance the variance due to observers and the variance due to the daily variation in teacher behavior. Such a design is referred to as $2 \times 4 \times 4$ design, since there were two observers, four visits to classrooms, and four teachers.

Degrees of Freedom

An important concept is that of *degrees of freedom*, a concept derived originally from physics that has been used extensively in the area of experimental design. If a body can move only in a plane, it is said that the body has two degrees of freedom but is restrained on a third degree of freedom. However, in statistics, the term has assumed a rather different meaning, which is related to the other only by analogy. Possibly the simplest explanation of the concept is to think of degrees of freedom as independent comparisons, a concept that needs to be amplified at this point.

For the sake of explanation, let us consider a grossly oversimplified case in which two pupils X_1 and X_2 received one treatment (say in reading), while two other pupils X_3 and X_4 were exposed to another. The comparison that it is desired to make between the reading scores of these pupils would be between the two treatments and might be represented by the term

$$(X_1 + X_2) - (X_3 + X_4).$$

An estimate of the variation to be expected from sources other than treatment would be provided by the two additional independent comparisons

$$X_1 - X_2$$
$$\text{and } X_3 - X_4.$$

These comparisons are referred to as independent comparisons from each other and also from the previous comparison, because the numerical value of each one cannot be determined from the other two. However, if an attempt were made to introduce another comparison, such as $X_1 - X_4$, it becomes clear that this comparison can be calculated from the other three and therefore cannot be considered to be independent of them. The number of independent comparisons that can be made in a system of observations represents the number of degrees of freedom of that system.

Four observations permit three independent comparisons. Ten observations permit nine independent comparisons. In general terms, one may state that N observations permit $N - 1$ independent comparisons, which represent $N - 1$ degrees of freedom.

Now if the comparisons that it is desired to make are 3 in a particular experiment and there are 30 degrees of freedom in the system, this means that there are 30 − 3 comparisons or degrees of freedom that may be used for the estimation of error. Designs must always be such that after meaningful comparisons are listed, a sufficient number of other independent comparisons are available to estimate the experimental error with the required degree of precision. If an experiment is restricted to 10 observations, it would be meaningless to attempt to study nineteen comparisons, since this would provide no independent estimate of error.

There is a simple way of determining whether a series of comparisons are or are not independent. Consider the case of the comparison involving the four measures X_1, X_2, X_3, and X_4. We were concerned with the three comparisons that could be written out as follows:

$$X_1 - X_2$$
$$X_3 - X_4$$
$$(X_1 + X_2) - (X_3 + X_4)$$

The corresponding coefficients of these terms may be arranged as follows:

+1	−1	0	0
0	0	+1	−1
+1	+1	−1	−1

It should be noted that the sum of the coefficients in each row is zero, and also that the sum of the cross products of any two rows is also zero. Comparisons that satisfy these conditions are independent.

The Testing of Hypotheses

All that has been said in this chapter up to this point is based on the assumption that the experimenter has in mind certain clear-cut hypotheses to test. The methods of experimental design that have been discussed are such that they assist the experimenter in testing his hypotheses with the minimum amount of data for a given degree of precision, and sometimes they permit the formulation of generalizations that cover a wider range of circumstances than would be possible if the classical type of experimental design were used. It is common to formulate hypotheses in terms of a null hypothesis, which states simply that the particular difference in which the experimenter was interested and that is empirically determined by the data is a result of chance variations and of a magnitude that might be expected to occur. Since the testing of the hypothesis from the data involves the determination of the probability that such a difference or greater would occur by chance, there is a certain logic in stating all hypotheses in the null form. In the sense described, it is possible to test the null hypothesis, and if the chances are extremely small that the difference (or a larger difference) would have occurred by chance, we may be willing to accept the alternative possibility that the difference was a product of differences in treatment. The latter hypothesis cannot be proven in terms of the data, but it does become more and more plausible as the null hypothesis becomes less and less plausible.

The analysis of variance and related statistical techniques that are outside the scope of this book are the techniques used for testing the null hypothesis, and they always provide an estimate of the probability that a particular difference could have occurred as a result of variations produced by sampling. The probability value that must be reached before it is decided to reject the null hypothesis is a matter of judgment, but it will depend on the consequences of mak-

ing an error. In this connection, two types of error have been distinguished and have become known as *errors of the first kind* and *errors of the second kind*.

In the particular experiment comparing the effect of two methods of teaching reading, the results might be presented in the form of the statement, "The probability that the observed difference or a greater difference would occur as a result of chance variations alone is 0.1." This means that if the two treatments had no differential effect, the results of sampling would produce a difference this size or greater in 10 per cent of all experiments. The experimenter might say to himself that this probability is small, and hence it is reasonable to conclude that the observed difference is generated not by chance variations in the sample drawn but by differences in treatment. If more thorough experiments were carried out later, substantial evidence might be collected to show that this conclusion was wrong. If this were the case, the experimenter would have made an *error of the first kind*. In this type of error, the null hypothesis stating that there is no difference between treatments is rejected when it should have been accepted.

Errors of the second kind are exactly the opposite type of error. They represent the case where the experimenter accepts the null hypothesis when he should not have done so. It is not possible to say that these errors are more or less serious than errors of the first type, because everything in this respect depends upon the circumstances. Fortunately, in education, the result of committing either type of error is not likely to be catastrophic, but in experimentation in other fields an error of either type may on occasion result in the loss of human life. In the interpretation of experimental results, it is important to keep in mind the consequences of each one of these types of errors. If the penalties involved in committing one of these types of error are heavy, caution must be taken in arriving at a conclusion that may make these penalties take effect.

The Design in Relation to the Question Asked

The care that needs to be exercised in relating the design to the question asked is well illustrated in a study by Mitzel et al. (1953), which was developed to illustrate some of the problems that this may involve. The Mitzel study involved the use of data from a previous

study in which two observers made four visits apiece to each of four classroom teachers. From these observations, a climate index was derived. Thus thirty-two measures of this index were derived. Mitzel et al. noted that three major questions might be asked of these data, which may be stated as follows:

1. Did the two observers "detect differences among the four teachers' behaviors on the four particular occasions on which they were visited with respect to the . . . climate index?"
2. Is it reasonable to hypothesize on the basis of the data that, if other observers visited other teachers on other occasions, differences between teachers other than chance differences would be found?
3. If the two observers were assigned to observe other teachers in the population of teachers from which the four were drawn, would other than chance differences in the climate index be found?

The analysis involved in these three cases is considerably different. The particular question to be asked in this case depends much on what are the experimenter's future plans and what he intends to do with the results.

The first question is a rather trivial one. It is of interest only if we wish to know whether the same observers visiting the same teachers on different occasions would still be able to observe differences between those teachers. The answer tells us absolutely nothing about the differences that might be found if other teachers or other observers were used. The model used makes the assumption that our observations represent a sample of observations that these observers might have made with these particular teachers. The universe sampled is circumscribed completely by these four teachers and these two observers. In this case, the answer given to the question asked does not seem to be particularly useful. At the most, it gives some indication of the reproducibility of the results under the highly restricted conditions specified.

The second question provides a much more useful answer, if it can be answered. The experimenter is likely to be interested in predicting what would happen if other observers were used with other teachers. In order to answer this question the four teachers must be

considered as a sample of a universe from which other teacher samples could be drawn, and the observers as a sample of a universe of observers from which other observers might be drawn. We are interested in generalizing from our observers and teachers to other observers and teachers. If the results of the study are to be used by other research organizations, this type of generalization is necessary.

The answer to the third question is particularly useful if the study under discussion has been conducted as a preliminary to a more extended study using the same two observers. It indicates something about the results to be expected from these two observers.

Sampling and Problems of Generalization in the Design of Studies

The design of experimental studies and investigations in the behavioral sciences, as in the biological sciences, is intimately connected with the problem of sampling. The intention of the writer cannot be to provide the student with an adequate background in theory of sampling, so our purpose here can be only that of making the student sensitive to some of the problems in the area so that he can turn to more comprehensive works to learn about the details of their solution —at least insofar as these problems have a solution.

In the testing of almost any hypothesis by statistical means, an assumption is always made that the observations recorded represent a sample drawn from a defined universe by methods that do not introduce bias. Some of the factors that introduce bias into the drawing of a sample were already considered in the chapter on survey techniques, with particular reference to the problem of identifying a limited number of persons who are to be interviewed from an unidentified universe of persons. In this chapter, the more general problem of obtaining samples of specified universes will be discussed.

Suppose that in a large city school system the director of research decided to survey the reading abilities of children who had passed their ninth birthday but who had not yet reached the age of nine years and six months. In this school system, a few less than ten thousand school children fell within this age range and these were distributed among twenty-five schools. It was clear that the director of research could not test all of these children on the particular test to be used. Therefore he decided to test a sample and to use statistical methods for making inferences concerning the total population from

the scores derived from the sample. A member of the board of education immediately suggested that it would be administratively convenient to limit the testing to pupils in a single school, since these pupils could then be tested together in a single session. The director was quick to point out that it was a well-established fact that one could not justifiably make inferences from the reading performance in a particular school to the reading performance in all schools, since average scores on the particular test in earlier years had been shown to fluctuate substantially from school to school.

Thus it is clear that the accuracy of the inference made from the sample to the population will depend on the way in which the sample is selected. The suggestion of the member of the school board lacks merit because it introduces bias into the sample for the sake of administrative convenience. Whatever sampling procedure is used, it is absolutely essential that it does not include any systematic bias. The simplest method of obtaining a sample from a population is that of obtaining a random sample, which is simply a sample in which every case in the population has equal chance of being included. By definition, the sample deliberately derived from one school could not be considered as a random sample because cases from other schools would have no chance of being included in it. One way of obtaining such a sample would be to obtain a list of all such children, then number them consecutively, and then select from this list by means of a table of random numbers. Tables of the latter type can be obtained from libraries. In using such a table, it would be appropriate to start by taking the first four digits and selecting the child who had the number corresponding to these four digits. The investigator would then take the next four digits and select the child whose number corresponded to these digits, and so forth. Thus each child in the population identified would have equal opportunity for being included in the sample to be studied. Under such conditions, well-established procedures can be used for making inferences from the sample to the population, and it is justifiable to neglect the fact that the population is not unlimited in size as such procedures require. The fact that the latter assumption is not fulfilled is not likely to affect our inferences appreciably when the number of cases is as substantial as it is in the present case.

In this simple type of inquiry, it is presumed that the director of

research is interested in estimating the mean reading score of the defined population from the sample. In the case of the random sample that has no systematic bias, the best estimate of the population mean is the sample mean. It is of course expected that there will be a difference *e* (*e* for error) between the sample mean and the population mean, but since the method of sample introduces no systematic bias, if the inquiry were repeated with new samples one would expect *e* to be as often negative as positive.

In this very simple inquiry, much can be done to reduce the value of *e* to a minimum. If the investigation is efficiently designed, it will be possible to obtain an unbiased sample such that *e* is smaller than it would be with a random sample of the type discussed. In essence, what is done is to take steps to insure that the sample is as far as possible *representative* of the population sampled with respect to important characteristics that are related to reading. For instance, it is known that girls show a tendency to be better readers than grade school boys of the same age. Hence it would be desirable to insure that the sample included the same proportion of boys and girls as was included in the universe under consideration. Since neighborhood is also related to reading skill, it would also be desirable to insure that the schools were represented in the sample in the proportion to the actual enrollment of the particular age group under study. Thus the sample would be stratified, and by making the sample more and more closely representative of the population, the tendency would be for the error term *e* to be steadily reduced.

Let us now consider a slightly more complicated problem of design in order to illustrate the relationship of problems of design to problems of sampling. Suppose that the director of research had been asked to evaluate a remedial reading program. In this program, approximately one hundred children in the elementary grades were given special remedial training in reading each year. The problem may be stated in this way: "If the reading skills of the pupils are measured at the end of the year of special remedial training, what is the probability that a random sample of children, similarly selected but without training, would perform as well or better than the trained sample?" It is therefore necessary to estimate the reading characteristics (mean and standard deviation) of the population from which the remedial reading group was a sample, for the time at which the remedial group

finished their special training. We may hope, of course, that the director of research did not find himself in the position in which many of his colleagues had found themselves—that of having no way of determining just what population had been sampled in the first place and then trying to remedy the situation by inferring the population from the sample. What he should have done was to exercise the most careful control over the selection of the original sample, so that he would know just what population had been sampled in selecting the pupils for remedial work. Unless this had been done, the data collected at the end of the experimental year would have been of extremely limited value if not entirely worthless. The prevention of such common tragedies is a matter with which the student of education must become fully cognizant.

The director of research would do well to start by identifying the population to be sampled. One way of doing this might be to administer a reading test to all grade school children at the beginning of the year. In each grade (or possibly in each age group), the lowest 10 per cent or 5 per cent might be considered to be the population eligible for remedial training in reading. Another method of identifying the population might be to define it in terms of the cases recommended by teachers, but this latter criterion is likely to provide a highly variable population from year to year. Therefore, let us assume that the director of research identified his population in terms of a cut-off on a distribution of test scores. His next step would be to select a sample to which should be administered the remedial training, and, at the same time to select a second sample that would also be followed up but would be given no remedial training. The latter sample, referred to as a control sample, would be used for estimating the population characteristics at the end of the year with respect to reading skills so that a determination could be made of the probability that the trained group could be considered as a sample of the untrained population. The investigator will attempt to make both his experimental and his control group as representative as possible of the identified population.

In all sound experimental design, it is important to start by defining the population to be studied and then to establish methods for sampling that population that will maximize the information supplied. All too often the reverse is undertaken. The author is aware of a book

that describes the behavior of four cases of reading difficulty. These four cases are presented without any inkling of the nature of the population of which they may be considered to be a sample. The reader is thus left wondering about the inferences that can reasonably be made from the data provided by the sample of four. On the other hand, if it were known that these cases were every tenth case admitted to a reading clinic for children aged eight to twelve years in a large city public school system, it might have been possible to make certain statistical inferences from the behavior of the sample to the behavior of the population sampled.

This example illustrates the methodological error of studying a number of cases and then seeking a population of which the cases could reasonably be considered to be a sample. When such a population is believed to be found, an attempt is then made to draw inferences about the population from what is known about the sample. This procedure is quite unjustified as a basis of statistical inferences. At the best, one can make what may be termed judicious inferences, keeping in mind that these are based on the assumption that the population considered is identical with that actually sampled.

The reader may ask at this point, "Would it not have been as satisfactory if the director had taken a group who had passed through the remedial training program and matched them on the basis of initial reading test scores with a group that did not have special treatment?" This question implies that a comparison would then be made at the end of treatment between the treated and the nontreated group. This procedure represents common practice in educational research, and its weakness is not apparent at first sight. The difficulty with which it leaves the investigator is this:

If it were found on the basis of a statistical test applied to the final reading scores that there was only a small probability that the two groups of scores could be considered as samples from the same population of scores, one would be led to expect that, if the experiment were repeated with a new sample, similar results would be achieved. The difficulty, however, is that we do not know just what population should be sampled in order to achieve the same experimental results. Those included in the original remedial reading sample may have been simply cases that had been chosen because they appeared to be those that could be treated with a high degree of

success; or they might have been a group that for some unknown reason responded well to treatment. The reverse might also have occurred if negative results had been achieved, for the investigator would not know to what population the results could be generalized. It might also have happened that a group that responded particularly poorly had been selected. In the case of the positive results, it would have been shown that there was some likelihood that positive results could be achieved by remedial training, but it does not tell us with what group positive results are most likely to be achieved.

Individual Differences and Block Design

Research designs of the block type, which originated in the work of R.A. Fisher, are unsatisfactory in the way they handle the matter of individual differences. This may be explained by means of an example. Suppose that a study were being conducted to estimate the extent to which differences in pupil satisfaction in different classes could be associated with teacher differences. In this study four high school teachers were selected, who were each teaching four different classes of thirty pupils. Two of these teachers were judged to be the most intelligent in the particular school, and two were judged to be the least intelligent. The pupils in each class were also divided into the fifteen more intelligent and the fifteen less intelligent in terms of a well-known intelligence test. The experimenter had in mind the hypothesis that the ablest students derived satisfaction from the ablest teachers but not from the least able teachers, and that the reverse was true in the case of the least able students. Now the experimenter in this case was undoubtedly thinking in terms of a continuous distribution of the intelligence of teachers, and probably assumed that what happens in the case of the two extreme groups of teachers can be used as a basis of generalization to intermediate groups.

However, such a generalization is not justified. The relationships established with extreme groups may not represent two points on a linear continuum. For all one can tell, the responses of pupils to the intermediate teachers may be quite different from what it might be expected to be on the basis of that assumption. This situation can be remedied to some extent by including intermediate groupings, but the inclusion of more groupings adds greatly to the complexity of the design. Correlational analysis of the type that has been used tradi-

tionally for the study of individual differences has advantages over block designs in many studies in which human characteristics must be incorporated.

Brunswick's Representative Design

Particular attention to the problem of generalizing from experimental results has been paid by Brunswick (1947). This writer points out that thinking in psychology is still influenced largely by classical experimental design, in which an aspect of some phenomenon is isolated and then studied under laboratory conditions. Thus in psychophysics, the aspect of the phenomenon of visual acuity that has been most closely studied is the ability to perceive two closely situated points of lights as distinct points. The separation that such points must have before they are perceived as separate by a particular individual would be considered a measure of the visual acuity of that individual. In further classical types of experimentation with this problem, the relationships of numerous conditions to visual acuity, as thus defined, have been studied. The typical and nineteenth-century-approved classical design involved holding all conditions constant except the one that the experimenter was manipulating.

The major change that has occurred in the classical type of design is the one due to the impetus of R.A. Fisher, which has been discussed briefly in this chapter. However, it may not provide results that are any more generalizable than the results of classical experiments.

Brunswick points out that these designs have one central weakness that has been disregarded. While they are usually planned with the purpose of including a sample of cases that is representative of a particular population or set of subpopulation, they fail to sample the variety of conditions to which it may be desired to generalize the results. In a great number of psychological experiments, the results found under one set of conditions may not be reproducible under other conditions. Indeed, it sometimes happens that one laboratory is unable to reproduce the results of another laboratory. Even in such a simple experiment as that previously described in which visual acuity is to be measured and studied by means of two points of light, it is doubtful whether the results are satisfactorily generalizable to other situations. If the purpose is to obtain results that can be generalized to other situations, the results may be disappointing. A

person who has relatively low visual acuity in the laboratory situation may do surprisingly well in other situations, for it is known that visual acuity is related to the general nature of the visual field, the intensity of surrounding illumination, the wave length of the light involved, the state of adaptation of the eye, and so forth. What Brunswick suggests is that we sample these conditions systematically in order to obtain results that, by and large, are applicable to these varied conditions.

Brunswick developed at least one example of a representative design that involved a problem of size constancy.* In this design, size constancy was measured under a great many different conditions, such as in a closed space like a room, outside the building, under different illuminations, etc. The purpose was to derive principles that could be applied under these varied conditions. Similar types of representative design are extremely difficult to undertake on matters of educational interest. For example, it would be valuable to be able to measure the expressed attitude of white children toward Negro children under varied conditions in order to predict related behavior under those conditions. Although the researcher might want to do this, the probability is that he would be able to measure such an attitude only under classroom conditions, and this would provide inadequate information for predicting what expressions and other evidences of attitude would occur under other conditions. Representative designs in the attitudinal area are rarely feasible.

The acute reader at this point may well ask why it is that a "representative" design is suggested. Would it not be simpler to list the extraneous variables that might affect the outcomes of an experiment and then incorporate them in a block design of the Fisherian type? When this can be done, it is of course the recommended procedure, but usually in research in the behavioral sciences the conditions that affect a particular phenomenon are so numerous that it is not feasible to incorporate them in a block design, for an unwieldy number of blocks would be involved. In most cases, it would also appear that these incidental conditions, though numerous, each contribute only a small effect, and probably too small to produce

* Size constancy is the tendency to see objects as being of a given size even though the distance between the object and the observer varies. Thus, a Cadillac appears to be a big car even though it is viewed at a distance of several hundred feet.

significant results in a feasible block design. In the face of this type of situation Brunswick suggests that a systematic effort be made to obtain representative samples of these conditions.

The concept of representative sampling that Brunswick has developed has come in for much criticism. One criticism that the reader will probably already have considered is that major advances have already been made in many areas in the behavioral sciences without resorting to the elaborate procedures that Brunswick's system demands. The psychology of learning is an example. Most of the important facts and principles of human learning that are discussed in typical textbooks were derived from laboratory experiments. For example, the principle that knowledge of results is an essential condition for learning was derived from a consideration of learning as it occurs in the laboratory and was demonstrated with simple laboratory experiments. Yet it seems to have wide application in the field of teaching.

A second criticism is that it results in the production of probabilistic laws—that is to say, laws that state only that there is a certain probability of a certain event happening as a result of a given set of conditions. Because Brunswick's system permits prediction over a wide range of situations, it is inevitably limited in the accuracy it can achieve. On the other hand, the more traditional approaches, because they aim ultimately at establishing *all* of the determinants of a particular event, have the ultimate aim of perfect prediction.

Summary

1. Courses in experimental design provide the student with a limited range of techniques for planning studies. However, a well-planned study can still pertain only to trivia.

2. Statistical methods serve two main purposes. One of these is the testing of hypotheses. The other is the summarization of data.

3. Well-designed studies have certain characteristics. They are free from bias, which may be introduced in various ways, some of which are not easily discerned. They must provide some satisfactory way of estimating error. They must insure sufficient precision to provide answers to the questions that are asked. The design must also be such that it yields as much information as possible from the number of observations that are made.

4. The term "control" is used in a number of distinct senses. It may

refer to the control of conditions that may interfere with the outcome of a study; it may refer to the control of the crucial variable that is being studied; or it may refer to the use of control groups or control observations.

5. Replication is introduced in order to increase the precision of a study, that is to say to increase the accuracy with which the main effects can be estimated.

6. An alternative to replication for the purposes of increasing the precision of a study is to control some of the sources of error. The traditional method of doing this was by a matching procedure, which has now largely been replaced by statistical methods.

7. The multifactor design is becoming more and more commonly used. Not only does this design permit the estimation of the effect of more than one variable, but it also permits the estimation of the effects produced by the interaction of these factors.

8. Sometimes designs are adopted that will not permit the estimation of all of the interaction effects. In such designs some interaction effects are said to be confounded with other interaction effects. Such designs commonly are used when interest is centered on the main effects, and they result in economies in the collection of data. An example of such a design is the famous Latin square.

9. The extent to which results have to be clear-cut in order to accept or reject a particular hypothesis depends on the consequences that follow if a mistake is made. If the consequences are serious, it is necessary for the data to present much more clear-cut results than when the consequences are relatively unimportant.

10. In checking the design of a study, it is important for the researcher to be sure that it will answer the questions that are asked, not merely closely related questions.

11. Brunswick has pointed out that experiments should be representative with respect to the situations they sample. However, the extent to which such a procedure should be adopted is controversial.

Data-Processing and Reporting | 15

Data-Processing

The plan for the processing of the data should be made at the time when the study is designed. By this is meant the time when the final plan is evolved. Of course, some preliminary studies have to be undertaken to insure that the enterprise is feasible. This is a more important matter than it may seem to be on the surface, and perhaps its importance may be brought home by citing an example.

A student once approached the writer with a proposed study of the effectiveness of two methods of teaching typing. The design of the experiment was a familiar one, with several pairs of matched groups assigned to the two methods. At regular intervals throughout the training program tests were taken by the students. These tests required rather prolonged periods of typing, lasting for as much as an hour each. The experiment was to be conducted over two semesters. During the conference with the student on this matter, a rapid computation was made of the volume of data to be collected and the time it would take to derive the scores that would be subjected to analysis.

As nearly as could be determined, the work would have taken the student about six months of his full-time attention. Also, the data consisted of sheets of typing and were such that it hardly seemed possible to design a device that would result in the quick scoring of the material. It would be an unreasonable use of the student's time to spend six months in clerical work, since this period could be much better spent in training related to his professional goals.

Various devices may be used to facilitate the derivation of scores from the raw data. One of these is to use a stencil scoring device if tests are in question. Never conduct research in such a way that the answers to a test are marked in a booklet unless it is absolutely essential to do it in this way. A separate answer sheet is a compact method of recording raw data. If the scores are to be converted to standard scores, then it is sometimes convenient to print the conversion table right on the answer sheet. If possible, the researcher should avoid having scores recorded on both sides of the answer sheet, since it is inconvenient to transcribe these scores onto rosters. Sometimes the separate answer sheet should not be used. Whenever speeded tasks of simple functions are involved, the operation of finding the appropriate place on the answer sheet and marking it may contribute more to the variance of the test than the function it is desired to measure. In such cases, it is obviously desirable to avoid the use of a separate recording system. What can be done in such a case is to print the problems right on the answer sheet, above or beside the place where the answer is to be recorded.

An alternative to the answer sheet is a version of the IBM punched card. Such cards are familiar enough to the reader through their several common uses—as checks, as bills, and so forth. They may be printed so that they have spaces on them similar to those found on answer sheets. The cards are then marked with a soft pencil, just as would be answer sheets, but of course they cannot be scored with the usual test-scoring machine. Instead they are run through a machine that converts the marks to punched holes. A computer can then be used to derive scores for each card.

Test-scoring machines can be adapted to the analysis of all kinds of data. At the present time the common type of scoring machine is one built by International Business Machines. This machine not only scores but also is fitted with an item-counter device. This device per-

mits the counting of the number of answer sheets that are marked in a particular position. It is thus possible to run a number of answer sheets through the machine and to determine the number that chose the first answer to the first problem, the second answer to the first problem, and so forth. Sometimes sociological data may be collected using an IBM answer sheet. If this is done, the scoring machine determines the number of cases with incomes in various brackets, or the number of persons in each age category, or whatever other breakdowns are recorded on the answer sheets.

The scoring machine just discussed will probably be considered obsolete within a few years. International Business Machines already has an experimental model of a new machine of a greatly improved type. A new concept of answer-sheet scoring has been introduced by Lindquist, who has built a machine at the State University of Iowa that is capable of performing most of the scoring required in the United States at this time. This work can be performed at a highly competitive price. The advantages and disadvantages of a single and centralized scoring service are not known at the present time but still need to be explored. The reasonableness of the scoring charges of this monster machine will make many research workers consider undertaking larger projects than they might otherwise consider. A picture of the Iowa scoring machine is shown in Figure XII.

If verbal material is to be quantified, it should be collected through the use of directions that require the individual to produce only a limited volume of words, unless of course this is inconsistent with the purposes of the research. If too much emphasis is placed on quantity, the product may be time-consuming to score. If much written material is called for, it should be divided into brief sections by means of appropriate directions. Sometimes instructions can be given to subjects to ask them to prepare material in outline form rather than as continuous prose.

Occasionally an experiment can be redesigned so that the quantity of data involved becomes more parsimonious. For example, one piece of apparatus designed for studying the problem-solving process required the subject to obtain information about the working of the machine before the problem itself could be solved. The apparatus was built so that the information-gathering procedure of the subject was recorded in detail, and from such records it was hoped that the

relationship between different information-gathering procedures and success in solving the problem could be studied. Such studies would have involved the analysis of vast quantities of material recorded on tape. This could have been avoided by comparing the problem-solving ability of groups trained in different techniques of gathering information. In the latter case, the data-gathering procedure would not have

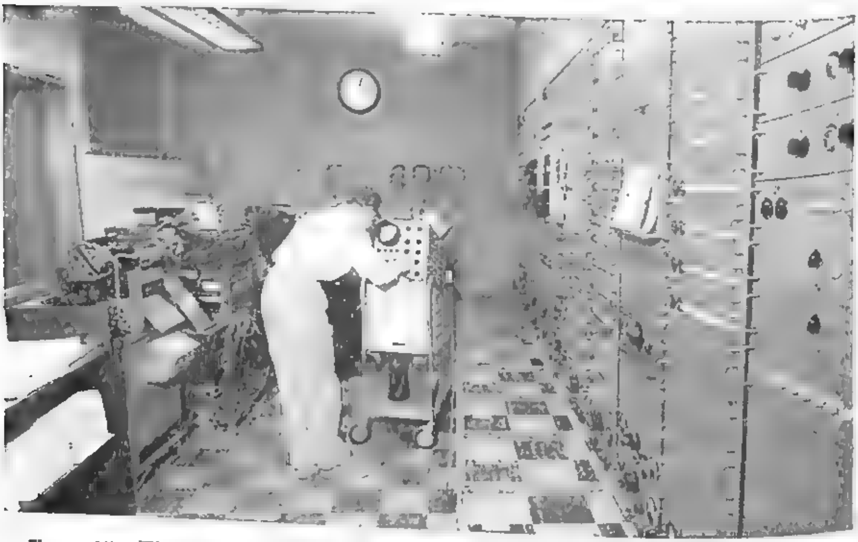

Figure XII. The Iowa scoring machine. This large scoring machine developed by Dr. E.F. Lindquist and his associates at the State University of Iowa is capable of performing numerous scoring operations at high speed. It has the capacity for scoring most of the testing programs administered in this country at the present time, and its services are available to anybody willing to pay scoring charges. Photo by courtesy of the Iowa Testing Programs, State University of Iowa.

to be recorded. The only information needed would be whether each subject did or did not solve the problem.

Some ingenious investigators have used a plan to reduce the clerical work involved in the handling of data, but one that is not endorsed here. The procedure is simply that of requiring the subjects to undertake the clerical work. Where responses are to be coded, the subjects perform the coding; where the tests are to be scored, the subjects score the tests. This is an undesirable practice, for two reasons. First, it introduces sources of error variance over which the

researcher may have no control. This is to some extent true even when a simplifying device such as an answer sheet is used. At least some error variance is introduced through errors in marking the answer sheet, but this becomes particularly pronounced when a speeded function is involved. If a complicated recording procedure is used, substantial errors may be introduced by the process. Second, a problem of ethics is involved. The researcher may have some justification in asking for the time of persons for the purpose of advancing knowledge, but he must respect their time and ask them to do only what is essential. The researcher should not be guilty of exploitation. Of course the issue does not arise if the subjects are paid, except that it may be much more efficient to employ a few well-supervised, trained clerks than a large number of untrained persons.

What to do about missing values is a particularly perplexing problem to which there is no completely satisfactory solution. In studies involving the analysis of correlation coefficients, a missing value in a table of raw data is of little concern. It does not matter much whether the coefficients in a table are based on slightly different numbers of cases. In factor analysis and in many other mathematical methods that are used for structuring data, slight variations in the number of cases from coefficient to coefficient are of little consequence.

On the other hand, when block designs are being used as the basis for an experimental design or as a basis of any other type of research, the problem of missing values becomes acute, since the computational methods that have been developed and that form the basis of tests of significance require the use of all cell entries. If certain cells are disregarded, the net result is to introduce an unknown amount of bias into the test of significance. There is no point in applying a test under these conditions, since it does not yield any kind of answer to the question that is posed.

At one time it was commonly suggested that mean values be substituted for the values of missing observations. The argument was that the measures were presumed to be normally distributed, and in such a case the class interval that includes the mean includes also the most frequently occurring values. Thus the insertion of the mean is an attempt to substitute the most probable value for the missing one.

Another approach is to compute expected values for those missing

from the other values provided by the data. Through the computation of regression equations, it may be possible to provide a least squares solution to this problem. However, this procedure is likely to produce more internal consistency in the resulting data than it would otherwise have. It will also bias tests of significance to an unknown degree. This problem has been worked upon, and proximate solutions that attempt to eliminate bias in tests of significance have been developed for many of the commoner block designs. The reader is referred to Cochran and Cox (1950), who have provided excellent accounts of the usefulness of these solutions and who have summarized research on this matter.

Another problem that sometimes arises is that of whether to discard certain observations that for one reason or another fall far outside the range of the remainder of the observations. Such discards must not be made after a preliminary inspection of the data has shown that the discarding of certain observations would make the data more in accordance with expectation. Discarding *must* take place before the significance of the data has been examined, for if this rule is not rigorously observed, the tests of significance that are applied will probably be biased. On the other hand, if this practice is observed, there is no reason why the researcher should not set up rules for discarding observations. These rules must apply to all observations, never only to certain groups. An example may illustrate this point.

In the last few years substantial interest has been shown in eyelid-conditioning, not because of any intrinsic interest in the phenomenon itself but because it can be used for studying a wide range of problems of learning. Eyelid responses are usually produced by a slight puff of air directed on the eyeball. However, only some of the responses appear to be true reflexes. On the contrary, some have a latency period, a delay in occurring, that makes many observers classify them as voluntary responses. In experiments that use eyelid conditions as the medium through which knowledge is acquired, it is customary to discard from the data those responses that occur more than a certain given time after the stimulus is applied. Different delays are permitted by different experimenters, but each experimenter sticks rigidly to the rules that he has set for discarding observations.

The researcher should always be on guard lest the procedure established for discarding observations does not by some means affect tests of significance that are later applied. This can happen in many ways, but the basic effect is always produced by there being a greater number of discards in one group than in another. Reasons why this may occur should always be carefully scrutinized, but even when the greatest care is taken to avoid any bias of this kind, there is always a faint possibility that a bias may have been introduced. The best rule to follow is to avoid conducting studies that involve the discarding of observations.

Observations may be recorded on rosters or on cards with numbered spaces. The writer's preference is for the latter system, since it provides greater flexibility and facilitates certain operations with data, such as the separation of groups of cases on which it may be desired to conduct special studies. The roster method of recording is a highly inflexible one, and even the correction of errors on rosters may present difficulties.

It is particularly important to check the accuracy with which all entries are made. The procedure is such a simple one that it often gives the false impression that it is just not possible to make errors on such a straightforward copying task. One very common type of error is the transposition of digits, such as occurs when a number is correctly read as 51 but incorrectly recorded as 15. Another source of error is the recording of digits in incorrect boxes on the cards or on the roster. All recordings must be checked with the most scrupulous care in order to catch such errors, for they may seriously affect the conclusions drawn from the data.

The processing of data presents certain problems that must now be considered. An important consideration in the data-processing procedure is that the scientist should *know* his own data. Unless there is a close personal contact between the researcher and his data, many important findings will never be made. Limitations may remain unnoticed unless close contact with the data is maintained throughout the processing procedure. For these reasons, there is at least some wisdom in performing a part of the data-processing by hand methods. This is no problem in the case of the student who is conducting research to fulfill the requirement of a master's degree, since the quantity of data is relatively small, and in any case it is probable that he

will process all of his data himself. On the other hand, if the quantity of data is large, it will be necessary to process all of it by machine methods, and in this eventuality it is difficult for the researcher to come to know his data as well as he should.

The student should be warned against the incorrect use of information derived from data. One such use is found in the researcher who gets to know his data well in order that he may derive from it hypotheses to be tested later by means of statistical tests. It should be remembered that statistical tests of hypotheses are not designed to test hypotheses derived from the data themselves. If such tests are applied to these hypotheses, they will produce answers that are biased.

The problem is perhaps better understood by considering an actual example. A research worker studying differences between delinquents and nondelinquents finds negligible differences between the two groups in all of the variables where he had planned to test the difference. However, a close scrutiny of the data reveals that the blue-eyed children who were unusually tall for their age showed a high incidence of delinquency, and this is advanced as one of the major conclusions of the study. The error made by the research worker in this case is that if one were to compare the two groups on a large number of characteristics, it would certainly happen that *in this sample* some combination of characteristics would be found that just happened to differentiate the two groups. There would be no reason for believing that the results would be repeated in a new sample.

In the above illustration, what the student should do is to list, during the planning stages of the study, all of the *reasonable* hypotheses that he proposes to study. His data should be collected for the purpose of testing these hypotheses and no others. All subsequently developed hypotheses squeezed out of the data would be subject to the criticism that they are not firmly rooted in the theory on which the study is based, and any apparent positive results would probably be the result of chance peculiarities of the particular sample of data.

The Use of Data-Processing Machines

The processing of mass data will undoubtedly be undertaken by punched-card methods or by other methods that use electronic com-

puters. If the graduate student of education can take a course in the use of punched-card equipment and can process his data as a part of the requirements of the course, then ideal circumstances are provided. All too often this is not the case, and the researcher can only specify for the machine operator the operations that are to be performed. Often more useful knowledge will emerge from intimate knowledge of the data themselves than from the highly complex analyses that are performed according to specifications.

Services now exist for the performance of extremely complex analyses of data. Modern high-speed electronic computers have been programmed so that they can perform routinely certain types of factor analyses and other types of multivariate analysis. In terms of the time it would take to perform similar operations by hand, the charges for these computations are modest. However, if it is desired to obtain a factor analysis on a twenty-variable matrix that will yield unrotated factors, the researcher at the present time may expect to pay at least two hundred dollars for this service. Simpler operations such as item analyses may be obtained at considerably cheaper rates, since they involve less valuable machines. A large data-processing machine is shown in Figure XIII.

After data have been machine-processed, an inspection should be made in order to catch certain errors that may be evident. For example, if the data have been reduced to distributions of scores, a check should be made to determine whether any scores reported are greater than the maximum possible score. Tables of correlation coefficients should be inspected for consistent signs. It would clearly be suspicious if a test correlated positively with one vocabulary test and negatively with another. Exceptionally large or unexpectedly small values for correlation coefficients also indicate that errors may have been introduced into the computational process. If checks have been built into the machine-processing procedure, then the inspection process is much less dependent on the ingenuity of the person who performs this activity.

If hand computations are undertaken with desk calculators, it is wise for the student to investigate whether short-cut methods have been developed for performing the desired operation. Most courses on statistical methodology in schools of education devote little time to problems of computation, since the emphasis is on understanding

and interpretation, and neither do most commonly used textbooks help in this matter. Here the advice of a statistician is important and can save the student much unnecessary labor.

Most of the graduate student's data are likely to be processed by means of desk calculators. As far as possible, checks should be built into the data-processing procedure so that the entire operation does

Figure XIII. A large data-processing machine. This large-scale IBM 704 Electronic Data Processing System, designed primarily for scientific and engineering computations, executes most operations at a rate of 40,000 per second. The machine's magnetic core memory stores instructions and problem data, any of which can be located and put to use in millionths of a second. Magnetic tapes and drums may be used for additional storage capacity. Photo by courtesy of IBM Corp.

not have to be repeated. When such checks cannot be built in, one should recompute all statistics, preferably by a different method and on a different machine, to insure that the final results are correct.

The student should be warned against a practice that is sometimes observed. This involves the checking of only those statistics that do not appear to be in line with expectation or with other values that have been computed. This partial checking introduces biases into the

apparent results and makes them appear to conform to expectation more closely than would accurately checked results. The reason for this is that a complete recheck of all results often changes some of the results that conformed closely with expectation, so that these show greater departures from expectation. The numerical results of a study must be reported with the highest accuracy that it is possible to achieve.

All work that involves computation must be carefully labeled. Most inexperienced researchers have the feeling when they are working on data that they will easily remember them when they come back to them at a later date. When they return to the data in a few weeks or months, they invariably find that they cannot identify what was done, and much time is lost in determining just what the computations signify. Often data remain permanently lost because they cannot be properly identified.

The data and the computations based on them should always be retained for a considerable time after the final report on the study has been published. This is necessary because a critical reader may desire to check other hypotheses with the data, or may wish to check the same hypotheses by different methods or after certain corrections have been introduced. It is embarrassing when some person questions the validity of the analysis and one is unable to produce the basic data so that the matter can be checked.

The Processing of Qualitative Data

Many of the data collected in educational research are qualitative. These present special difficulties when they are to be processed. In the early stages of acquiring information about a phenomenon, no attempt may be made to process carefully the facts that are collected. Freud's classic observations on the behavior of disturbed patients are examples of qualitative data collected and examined for the purpose of developing hypotheses, and the conclusions that he drew guided the research work of subsequent generations of psychologists. Early explorations are usually made in this way, but the time must ultimately come when such observations must be analyzed systematically. The mere inspection of data without the aid of systematic analysis is a hazardous process, and there is always danger that the researcher will dream into his data elements he wishes to see there

that do not really exist. For this reason every effort must be made to reduce such data to a form where they can be analyzed by appropriate methods. In this way personal prejudice can be eliminated from the interpretation of the material.

As a first step in the analysis of qualitative data, it is necessary to code the facts that are involved. This means simply that a number must be assigned to each class of fact. Thus, if the cumulative records of children are to be studied, it may have been determined that perhaps eighty items of information are to be coded. Those concerned with the coding operation might be asked to code all items of information on a sheet, a section of which might be as follows:

42. Progress through school	0 = never held back a grade 1 = held back one grade 2 = held back two or more grades
43. Absenteeism from all causes (average days per year over period of record)	0 = 5 days or less 1 = 6–10 days 2 = 11–15 days 3 = 16 or more days
44. Speech	1 = no speech difficulties reported 2 = speech difficulties but no action taken 3 = speech difficulties and remedial work started

Through such a code sheet the qualitative information obtained in the cumulative record is converted into a set of numbers, which are then used for the analysis. Sometimes the numbers are entered directly onto the code sheet, if it is quite brief. If it is long, the code numbers are often entered on a separate sheet or card. The code numbers on the cards may then be punched into other cards if the data are to be analyzed by machine, or the cards may be sorted by hand if a hand analysis is to be made.

Certain precautions need to be taken in adopting this type of procedure. First, it is necessary to conduct a trial run in order to determine whether the coding system involves any ambiguities. It might then be found that, for example, severe weather one year had

closed the school for ten days. Since this eventuality had not been taken care of in the coding system, an additional rule would have to be introduced into the coding of item 43, namely that absenteeism is not to include days lost through the closing of schools. Additional rules of this kind almost always have to be introduced when a coding system is tried out.

Any set of rules established for the purpose of quantifying qualitative material should be tried out further by submitting the material to different coders in order to determine whether the rules can be applied with consistency by different workers. This tryout helps to establish the error resulting from differences in the judgment of different persons. The tryout may also result in the development of methods that eliminate these discrepancies. The reliability of such procedures should always be given in the report of the study, unless the procedures are such that it is clearly evident that *careful* independent workers can produce independent results. The latter might be true where the entire process of quantification required only the counting of the number of words written in documents produced by subjects.

Errors are commonly made in clerical procedures that involve the coding of data. These are usually referred to as errors of carelessness, but in actual fact they are probably not so much due to the failure of the clerk to be conscientious as they are a product of the immensely boring nature of much of this work. The upshot is that it is essential for all clerical work of this kind to be rechecked independently by another clerk. It may not be necessary to lay down the standard that only perfect agreement is to be accepted, since this may be unrealistic, but some standard should be established. Scores that do not agree within one point, or two, or three, or whatever is considered reasonable, must be redetermined, preferably by another person. Clerks must work independently, otherwise this check would be quite meaningless.

WRITING THE REPORT

Excellent suggestions concerning the general format of the research report can be found in a guide prepared by the American Psychological Association for the convenience of contributors to its journals. The convenience of the style suggested in this report to both

the publisher and the printer has resulted in their being adopted by a wide range of organizations engaged in the publication of technical literature in related fields. The style is just as appropriate to the publication of educational research as it is to that of the type of research found in the journals of the Association. The reader might do well to become familiar with this style, which recommends treatments for such matters as reference sources, the use of footnotes and quotations, the handling of tables and figures, and general matters of usage. The document under discussion also makes recommendations concerning the way in which the manuscript should be typed and prepared for the printer. The reader should become familiar with such matters by studying this document before he engages in report-writing.

Beyond the initial words of advice given in the previous paragraph, the writer finds himself treading on uncertain ground. He can recall journal editors who marked the margin of some of his most prized paragraphs with the words "not clear." After a few such experiences, he wonders whether any writer should ever try to give advice to another on how to present his ideas. Nevertheless, he has been asked so many times by students and researchers to advise on problems of presenting research that he realizes that the student also will demand a section in this book on the same topic. The section that follows repeats the advice that he has given to graduate students on many occasions, and also repeats the criticism he has made of these students' dissertations. Whether such advice as will be given can improve the student's writing is a matter about which the writer cannot even speculate.

It is common for a dissertation to be written in at least two forms. The initial form presents the material to an examining committee. The second form, if written, is a condensed version that presents the material for publication. The initial version must be written, like any other piece of writing, with the nature of the specific audience well in mind. It is to the student's advantage that he know personally those who constitute this audience and be able to write specifically for them. To some extent, he should write with their expectations in mind. If one of them is likley to ask the student to relate his findings to some particular theory, then he should be sure to do this.

In the case of writing for formal publication, the problem is much

more difficult. The student would do well to start by reading other articles in the journal that is being considered as a place of publication. From this overview, he should arrive at judgments concerning the nature of the articles that the board of editors favors, and also their length and organization. Boards of editors, like any other groups, have personal preferences, and these must be taken into account since they may be the deciding factors in determining acceptance or rejection of the student's product. In the ultimate analysis, publication is achieved because the writer is able to arouse an affective response in a reader.

If the report is to be presented as a doctoral dissertation, it will be necessary to peruse the rules and regulations regarding the writing of such a document. Most graduate schools publish a list of special requirements in this regard. The doctoral dissertation presents fewer restrictions on space than are imposed by conditions of journal publication, but this is no reason for the student to engage in expansiveness in the development of his thesis. There was a time when major institutions required that the doctoral dissertation be published at the student's expense, and this provided an excellent incentive for brevity, but the modern practice of requiring only microfilm reproduction does not have a corresponding inhibiting effect. Similar rules have also been established for master's theses.

The Introductory Sections

The introductory sections usually begin with a statement of the problem. The writer feels that in the reporting of most research studies at least the general nature of the problem to be investigated should be found in the very first paragraph of the report. The statement may not be in a full and precise form at this stage, since it may first be necessary to introduce the reader to a number of terms and concepts before the problem can be accurately set forth, but nevertheless there should be a statement of the problem, even if only in a general form.

The introduction must also provide an appropriate theoretical orientation for the reader. This may involve a history of the problem and a review of related studies. In some cases, the theoretical framework of the problem may be so familiar to those who are likely to read the article that it may be quite unnecessary to state it except in general terms. For example, a student working on the problem of

reinforced learning would obviously not review reinforcement theory, which has been described fully in so many other sources. On the other hand, if the research is concerned with a theory with which the reader is unlikely to be familiar, it is essential that the theory and its background be outlined in the introductory section. If the theory is the researcher's own, it is desirable that it be fully presented in terms of the procedures described in earlier chapters of this book.

In the preparation and execution of a research, extensive work is often undertaken on the review of previous studies in the area. If it is done by a senior research worker with broad experience in the field, this may constitute a major contribution in itself. When substantial effort has been devoted to this phase of the undertaking, it is possible that a separate article may ultimately be prepared and published to cover the outcomes of this activity. Such contributions may form an immensely valuable contribution to the professional literature; but when such a separate contribution is made, it is necessary in the research report itself only to refer briefly to this article and to list it as a major reference, giving merely the major outcomes of this activity.

The review of the literature should lead up to the full and complete statement of the problem. If the introduction gives or implies the statement of a theory, as it should, the problem should be stated as a deduction or consequence of the theory. Earlier in the introduction, the student should have defined all the terms needed for understanding both the theory and the statement of the problem. By the end of the introductory section, the reader should be fully prepared for understanding the explanation in subsequent sections of how the problem was solved.

The Description of the Procedure

The vital importance of the section that describes the procedure or method is often not appreciated by the novice in scientific research. The criterion of a well-written description of the procedure or method used is whether it provides sufficient detail for another researcher to reproduce the study. Too often the experimenter writes up his work only to find that insufficient detail is given for another even to begin to reproduce it. In the behavioral sciences, the writer faces real difficulties in deciding what are and what are not the important details to report in describing his procedure. It is clearly quite impos-

sible to detail all of the conditions related to the undertaking of a study. For example, in describing an experiment, is it relevant to report that the experimenter was a woman, or that she was a blonde, or that she was born in Germany? The present writer knows of one study in which it was relevant that the experimenter was a woman, and the results probably could not be reproduced without a control over that factor. However, he does not know whether any experiment has been reported in which it was relevant that the experiment was blonde or was born in Germany. In any event, the fact that such a factor was important in one study does not mean that it would necessarily be important in another study. The decision has to be made in each case concerning what is to be reported and what shall be omitted. The fact that this decision must be based largely on judgment reflects our lack of knowledge about behavioral phenomena.

The description of procedures should include a reproduction of verbal directions given to the subjects. Where these are lengthy, they may be relegated to an appendix, or a footnote may indicate where a complete set may be obtained. Minor differences in wording may have substantial effects on the outcomes of a study. Unfortunately, matters of intonation and emphasis cannot be accurately described, although these may have substantial effects on the outcome.

The description of apparatus is likely to be unsatisfactory unless the greatest care is taken. Since it is not usually possible to publish a blueprint, it is necessary to specify the essential details. However, the experimenter sometimes may not know what are the essential details. This statement may need some explanation.

The author was concerned some time ago with the replication of an experiment that involved apparatus. The piece of equipment specified was the Harvard tachistoscope, which is widely used in psychological laboratories, and this piece of equipment was readily available. The object to be seen through the tachistoscope was illustrated in the original article, and this was easily reproduced by a draftsman. However, after some work with the equipment, it became evident that a crucial feature of the entire arrangement was the size of the object presented. This had not been specified in the original article, but the results could be reproduced only when the object was a certain size. The original experimenter had been unaware that this was an essential aspect of his experiment and had failed to report it. Unless

there is a great deal of replication with variation, the experimenter is likely to be unaware of the essential characteristics of his apparatus.

One advantage of using standard apparatus can be seen when the problem of description arises. It simplifies matters greatly to be able to report that a Harvard tachistoscope or a Brush Model 392 amplifier was used, since this equipment can be duplicated by other experimenters. The home-grown type of equipment needs careful description.

Sometimes the research worker calibrates apparatus, in which case it is necessary to describe the method and technique used in calibration. Sometimes the equipment used in calibration is as complicated as the apparatus itself.

A common omission in studies of educational behavior is a failure to indicate just who was included and who was excluded from the study. This is the matter of specifying the sample that was included, or perhaps one should say what universe was sampled in selecting subjects for study. There is the same need for specifying the universe that is sampled when the objects are inanimate as when they are living. The student will realize that unless the researcher knows how his sample is drawn, he will not know to what his results can be generalized.

In describing the method used for selecting human subjects, sufficient information should be given so that the reader may know to what population the results may apply and also how he can obtain a similar sample and reproduce all the conditions of the study.

Reporting the Results and Stating the Conclusions

The results of a scientific study should usually be presented in a table for which there is some explanatory material; but, since many studies in education do not approach ideal standards, this method of reporting cannot always be attained. A distinction should be made between the results of the study and the interpretation of the results. By the results is usually meant the summarized data and the test that is applied to determine whether they are or are not consistent with the hypothesis they were designed to test. Usually in educational research some test of significance must be applied to the data in order to test the hypothesis. It is usual to describe this test in the results section of the report and to indicate the assumptions that

are made in its application. Usually these assumptions will be approximated rather than completely fulfilled, as when it is said that distributions of scores can be considered to be samples of normally distributed universes. Sometimes, although the conditions required by the test of the hypothesis are not completely satisfied, empirical studies carried out elsewhere have shown that these departures will not affect the outcome. The results section should also describe any special and unexpected events that occurred during experimentation, as when subjects were unable to complete the schedule because of illness or other causes. The treatment of missing values should also be discussed in the results section.

As far as possible, the table or tables presenting the results of a study should be self-explanatory and should not require extended reading of the text in order to understand them. On the other hand, the material in the text should point out the important aspects of the data and draw attention to the relevance of the results.

Just how much tabular data should be presented is always a matter of judgment. As a general rule, only those statistics that are crucial to the testing of a hypothesis should be presented. Detailed raw data rarely can find a place in a research report, except where they are of such unusual interest that their reproduction is definitely in the interest of science. If the data are complex and it is anticipated that others may wish to rework them or to use them for testing additional hypotheses, it may be worth while to have them recorded with the American Documentation Institute. For example, the writer can recall that one investigator known to him had computed all the correlations between 150 test items. This table of intercorrelations might be of considerable use to other researchers working in the field of factor analysis and related methodology, and such data should be made available.

A common error in the presentation of results is the division of the data into too many separate tables. Many research reports can be improved by the consolidation of tables into larger units.

Some comment should be made on a special problem of reporting results that occurs from time to time. The problem is that of what to do with experiments that do not yield anything that can ordinarily be reported as results. Reference is made here not to experiments that yield negative results, for the latter can usually be reported by

the procedures discussed, but to experiments where some technical hitch occurs that prevents the experiment from being carried through to its proper conclusion. These abortive efforts are not entirely useless in the information they provide. Indeed, if the problems they raise are never discussed in the literature, others will attempt similar experiments and end in similar difficulties. The difficulty of reporting such efforts stems from the understandable unwillingness of editors to accept articles about them. To the present writer, the way out of this dilemma is to report the results of abortive experiments in the introductory section of a further experiment that was successful. One may preface a successful experiment with an account of the various avenues and approaches that were explored before it could be undertaken. Such an account can be brief, but it should be sufficient to warn others about the limitations of the alternatives that were explored.

This does not mean, of course, that weaknesses in the approach revealed during the course of the study should not be noted. Sometimes it is necessary and desirable to admit that the main knowledge derived from an experiment is how to design a more conclusive study. It also happens quite frequently that a study designed as a crucial and conclusive experiment turns out to be, on further thought, ambiguous in its results because of the various ways in which they can be interpreted. There are many well-known cases of such experiments that were designed by famous experimenters. For example, there has long been argument as to whether it is possible to have nonreinforced learning. Experiments have been designed in which animals, usually rats, have been provided with opportunity for learning without any apparent reinforcement. However, it has usually been possible for the protagonists of reinforcement learning theory to point to the possible existence of hidden reinforcing conditions, which destroy the conclusiveness of the results.

The conclusions should state the extent to which the data are consistent or inconsistent with the hypothesis or hypotheses. They should be stated in the same order as the original hypotheses and should parallel the list of them. The statement of the conclusions should not be contaminated with implications or with other types of speculative discussion. There is no reason why they should not be stated in a fairly terse form.

A common error is made in drawing conclusions from research results. This error is seen in cases where an investigator collects data that reject a hypothesis and hence question the validity of the postulate on which the hypothesis is based. Under such circumstances, some investigators are inclined to turn around on themselves and to seek reasons why the experiment was really not a crucial test of the issue it was designed to settle. The situation indicates either that the investigator had become too attached personally to his own ideas or that the test of their validity was inadequate in the first place. If the latter was the case, the question may be raised as to why the experiment was ever conducted. If the experimenter changed his mind during the research and began to question its utility, then he should have stopped his work and certainly not published his results.

Writing the Implications and Discussion Section

The creative research worker will inevitably speculate on the implications of his study that extend beyond his immediate purposes. He will also want to communicate his thoughts on such matters to a wider public. True, nobody is ever likely to treasure these thoughts as much as the creator of them, but nevertheless some may be useful to subsequent research workers, and a few may even be real gems. The section of the report dealing with implications may be used quite appropriately for setting forth these thoughts.

It is important that the section on implications should be more than a splurge of personal notions. Whatever ideas are presented must be set forth in a well-organized form. Sometimes it is convenient to organize them around a few areas where the implications have special importance. For example, in one study of mechanical problem-solving with which the author is familiar, the implications were organized around two topics; namely, the selection of mechanical trouble-shooters and the training of trouble-shooters. Good organization will develop in the reader of the report a better appreciation of the importance of the writer's ideas than will a poorly organized section.

Brevity in the implications section is also a very desirable characteristic. Most readers have only limited appetite for the speculations of others. A lengthy section may produce boredom to the point of rejecting even the good ideas that are presented, while a discursive style may, at times, be extremely useful for driving home a point.

However, a certain degree of crisp conciseness should be aimed for here, as in other parts of the written report.

The section on implications is also the section in which it is appropriate to give some indication of the future direction of the program of research of which the report represents a part. Perhaps it may be well to remind the reader again at this point that if research is to be profitable to the greatest extent, then it must be programmatic. A research report should end, therefore, not with a note of finality but with some indications of the unfinished business that should be the next preoccupation of the researcher.

If the report has been introduced with the presentation of a theory that the research is designed to extend or modify, then the final section may well restate the theory in the light of the findings. This process may involve such radical changes in the original formulation that what is virtually a new theory has to be stated. Whenever the research results in the restatement of a theory, it follows that the research report should indicate how changes in the theory should modify current practices.

Use of Diagrams, Tables, and Figures

A common error in the writing of technical reports is the failure to use diagrams effectively. The writer can remember more than one instance when he has had to wade through ten or more pages describing a complicated piece of apparatus when a simple diagram and a page of description would have sufficed. The writer also suspects that some readers are quite unable to translate verbal descriptions into visualizations of the equipment described. The medium used for communicating should be appropriate to the material to be communicated.

If diagrams of apparatus are given, and they are necessary if any apparatus has been used in the study, it is most desirable that they be prepared by an artist. This is not necessarily as expensive a procedure as it sounds. If an artist is provided with a good sketch, he is likely to produce a finished diagram with considerable speed. The writer has had many such made for approximately ten dollars each.

The artist or draftsman will have to be informed of the size to which a diagram has to be reduced. Usually, he will draw it on a larger scale and his drawing will be reduced photographically.

Figures and graphs should be presented in such a way that they are self-explanatory. The headings and captions to figures and graphs should provide all of the explanation that is needed. Discussion of what the table or graph demonstrates in relation to the hypotheses can be appropriately included in the text.

Typed manuscripts do not usually have an index, but if a dissertation or other research report is published as a monograph of substantial size, then an index is a most important feature. It is prepared during the page-proof stage, when final page numbers have been assigned to each portion of the text. A simple procedure for making the index is for the person who is reading the page proofs to have at hand a quantity of standard index cards or slips of paper measuring three inches by five inches or smaller. Each time a technical term or name is encountered in reading the text, it is noted, together with the page number, on a slip of paper, which is then deposited in a box. At the conclusion of reading the proofs, the slips are put in order to form an index.

In any larger work, an index is an invaluable aid to the reader and should be prepared with care and thoroughness. Its usefulness depends to a great extent on the appropriateness of the classification of concepts that it uses.

Other Points on Organizing the Research Report

When a research report is of such a length that it requires organization into chapters, it is most desirable to provide the reader with certain devices that will enable him to keep track of the argument and to find his way around in the mass of material. This can be done in several ways.

First, it is desirable to provide chapter summaries. These should help the reader to organize his thoughts by going over the high lights of what he has read and the conclusions and arguments presented. The summary should be strictly a summary; it should not include new material that happened to occur to the researcher after the report was written. It may well be organized into a series of numbered paragraphs, and these should be written in a concise form.

Second, a system of paragraph and section headings may well be adopted. Indeed, some writers like to do their work by preparing a list of headings and then writing the sections and paragraphs in any order, working at any one time on those where they feel that their

thinking has reached the point of maturity and where an organized statement can be put down on paper. Some writers prefer to use a system of major headings and minor headings, in addition to chapter headings, but this can be done only where the material lends itself to this type of organization.

Third, a good table of contents is a most desirable guide for the reader. Where paragraph and section headings are used, these should be listed in the table of contents.

Finally, brief mention must be made of style of writing in the report. He who tries to advise another on questions of style is treading on uncertain ground. When one sees how often literary critics have been wrong in predicting the acceptability of the works of writers, one realizes how unreasonably prejudiced one may be in one's preferences for style. Also, a person's style is dear to his heart, and suggestions that it be changed or even that it be criticized may arouse ire. Therefore this writer, acting with a certain sense of self-preservation, will at this time point out only certain common features of technical writing that detract from its value in communication.

First, there is the error of using too difficult a vocabulary level. A writer should not select a word just because it is appropriate and because *he* knows the meaning of it. A necessary condition for the use of the word is that the reader also knows its meaning. When unfamiliar words are introduced, the writer must remember that the reader will have to learn them. It is not sufficient that they be formally defined once and then used. This is like asking a person to learn a word by exposing him to it once. What one has to do in writing is to give the reader as much opportunity as is feasible to learn new terms. These must not only be defined but also be used in contexts where their sense can be inferred from the general meaning of the sentence in which the words are placed. The writer who introduces several unfamiliar terms and then fails to provide the reader with a learning experience is likely to find that most of his public does not read beyond the introduction.

A few technical writers have acquired the reputation of writing in a language that is familiar only to themselves. Such writers may have been careful to define their terms, but since these terms have not acquired general usage, readers have never learned them and much of the writing that uses this language is never carefully read.

Hence much of it is lost. For this reason, the reader should realize that new terms should be coined only when it is absolutely necessary to do so.

Just as unfamiliar words should not be used except where they are essential, so too is it desirable to avoid passing references to obscure theories with which the reader may not be familiar. If such a little-known theory must be mentioned, it is desirable to introduce it by presenting its main features. Such brief descriptions can be appropriately introduced as a part of the text. The nineteenth-century practice of using lengthy and elaborate footnotes to explain any obscure point in the text is one that has become less and less frequently used in scientific literature as time has passed.

Those who have not had previous experience in writing often manifest the error of unnecessary repetition. Now some repetition is necessary in most writing, and the old adage applies that the teacher should start by telling his audience what he is going to say, then he should say it, and finally he should say what he has said. A report, as much as a lecture, is a learning experience for the audience and a teaching experience from the point of view of the writer. Thus systematic repetition of the type described by the adage is a desirable feature of written presentation. The kind of repetition that should be avoided is that where the writer keeps on harping on a point or coming back to it.

A frequent error of style, particularly common in the literature of educational research, is that of writing out in extended detail facts that have been presented in a table in concise form. An example of this from a mythical report is given below:

> The table under consideration shows the percentage of correct answers to the arithmetic problems given by various categories of college students. It can be seen that freshmen, sophomores, juniors, and seniors obtained on the average 32, 34, 43, and 44 per cent of the problems correct. When the same group of freshmen is divided according to whether they came from type A, type B, or type C schools, the percentages are 29, 31, and 33. The corresponding figures for the sophomore group are 31, 32, 36; for the junior group, 41, 42, 45; and for the senior group, 43, 44, 45.

Drivel of this kind fulfills only the purpose of confusing the reader, who would have understood the data perfectly well if he had been left to examine a well-constructed table.

A similar stylistic fault is seen when a writer is attempting to explain a mathematical operation that he has performed on data and does this by giving an extended account of the arithmetic involved instead of providing a brief account of the algebra or of the general purpose of the operation. An example of this kind of error of presentation is the following:

> The totals for each one of the horizontal rows were squared and from the sum of these values was subtracted the square of the grand total divided by 500. The result of this operation was then divided by 6 and the dividend was entered in Table X. A similar arithmetical operation was then performed with the totals of the vertical rows, etc.

What the writer should have done was to state that he performed an analysis of variance according to customary procedures. If he wanted to explain further what he was doing, a brief algebraic explanation would have sufficed.

Final Publication

Most theses and dissertations do not achieve publication beyond that provided by microfilm services. This is usually not because the findings do not merit it but because the author does not take the necessary steps to incorporate his findings in the professional literature. Most doctoral dissertations from a well-established graduate school contain enough of consequence to provide at least one publication, and some may yield several. They would not have been accepted in partial fulfillment of the requirements for a degree if this had not been so. True, occasionally a dissertation of no consequence slips by through some misunderstanding, as when everybody involved had become committed to accepting it before its worth had been properly evaluated, but such cases are exceptions. Failure to publish the results of a doctoral project is almost always due to the fact that the student failed to take the steps necessary to achieve this goal. Often this is due to lack of motivation, since the achievement of the doctoral degree represents the achievement of a personal goal and publication has little to offer to the student. In addition, he may have already revised his product so many times that any further revision is seen as a most distasteful and repulsive task. However, the doctoral student, though aware of these blocks and impediments, should recognize

that he has to do more than consider personal gain in deciding whether to publish or not to publish. He also must consider that in the preparation of his dissertation he has occupied much professional time on the part of a faculty, and that he can repay society for this by making his findings part of the body of professional knowledge represented by published literature. If he does not do this, much of his time and the time of others will be lost, and later students may repeat his work without ever knowing that they are merely repeating what has already been done. Master's degree students have a lesser responsibility in this respect.

The most desirable place of publication is the professional journal that specializes in the field in which the student has worked. Many such journals publish without charge except for special materials such as tables and cuts. Less desirable as places of publication are those journals that require the authors to defray the cost. It is inevitable that the free-publication journals should be able to select the best contributions.

Most journal editors will provide considerable help in shaping an article so that it presents what is to be presented in the most effective way. Suggestions by editors concerning the revision of manuscripts should be given careful consideration. In such matters, the editor's experience is likely to provide a sounder basis of action than is that of the neophyte in the field. Editors are deeply concerned with making their journals into the best publications they can possibly produce.

New Methods of Distributing Technical Information

Twenty years ago it was a relatively simple matter to publish long articles that included substantial quantities of tabular material, but today lengthy scientific documents are extremely difficult to publish. A part of this change is due to the large increase in publication costs that has taken place over the period, but this is not the whole reason. An additional factor in the situation is the expansion of research in the behavioral and educational sciences, with the result that most journals receive many times as much material as they ever have space to publish. One partial attempt to solve this problem is for journals to add additional sections that are published at the expense of the author. A few writers have been attracted by this proposition, and especially by the feature that paid publication of this type results in

immediate publication and the usual long delay is eliminated. Nevertheless, the expenses of such early publication are high and beyond the economic circumstances of most young research workers, even if they are fortunate enough to have their article accepted.

Those who have thought about the problem of providing publication facilities, and hence of the problem of distributing scientific information, agree that traditional journal sources will become progressively even less adequate than they are at present. New methods of distributing scientific information must be found, and to some extent these are in the process of being developed. One attempt to develop a new technique in this connection is represented by the American Documentation Institute, which is commonly represented by the letters ADI. It is a private, nonprofit organization that provides a special type of service.

When a research worker is faced with the problem of publishing a study that includes, for example, large tables of correlation coefficients or variances and covariances, he may write his article so that only the analysis forms an integral part of the article. The original table of correlations or other material on which the analysis is based may be omitted, and a footnote is included in the article to indicate that these materials have been deposited with the American Documentation Institute and that the order number of the document is such-and-such. A reader of the article who wishes to refer to the basic data underlying the analysis may write to the American Documentation Institute and, for a small fee, may obtain either a photographic copy of the document or a reproduction on microfilm. Charges for this service are very moderate.

Since the American Documentation Institute provides only photo reproductions, it is necessary that all documents transmitted to them be suitable for this purpose. Typed material should be typed with a heavy black ribbon that is quite new, and there should be an absence of stray marks or messy erasures. It hardly needs to be pointed out that the name and location of the person depositing the document should be marked clearly on every document.

Documents deposited as a part of a study that is appearing in a journal should be submitted to the journal with the manuscript. This is done in order that the editor may better judge the entire manuscript and determine the relevance of the supplementary documents.

Brief mention may be made of a new organization within the Department of Defense, which may represent a common method of the future for distributing scientific information. This is the Armed Services Technical Information Agency, usually known as ASTIA, which represents an attempt to provide a pool of well-organized technical information derived from the research agencies of the Department. The Agency prepares an organized list of titles, referred to as the Title Announcement Bulletin, or TAB, and this is distributed to using agencies. At the same time the Agency prepares abstract cards, which are distributed to those who need additional information about particular titles, and some users are provided with complete sets of cards in specific areas if they so request. If the information on an abstract card interests a user, he may request a copy of the complete document. The latter is provided either as a full-size reproduction or in some form of micro reproduction. The copy is loaned and is returned when no further need exists for its use.

The ASTIA program seems to the writer to be in many ways a model for programs of the future. Perhaps it should be pointed out that Ohio State University took a step in this direction some years ago in preparing comprehensive bibliographies in the area of education and educational research, but the program was perhaps before its time and did not blossom into a document collection and reproduction service. When the ASTIA system is applied to civilian problems of disseminating information, it is suspected that the present system of journal publication will become obsolete. Many problems will have to be solved in this connection, but a new system is surely needed.

Summary

1. The plan for the processing of the data should be drawn up at the time when the study is designed.

2. Data should be collected in a form that is convenient for processing.

3. If apparatus is used that records mechanically the desired data, it should be arranged so that it delivers data in a concise and manageable form.

4. Procedures for quantifying data should be written out in detail and should be given a trial run prior to actual use in the study. The reliability of these procedures should be reported if they involve judgment.

5. The problem of handling missing values should be discussed with a

statistician if it arises. The discarding of observations may seriously bias the outcomes of a study unless the research worker knows how to handle this problem.

6. Clerical work should be checked for accuracy, since this may introduce large errors.

7. The research worker must avoid the error of deriving hypotheses from the data and then testing the hypotheses from the same data.

8. Large quantities of data require machine processing, which can be undertaken by organizations that have been established for this purpose.

9. The checking of statistical work must be complete. If the researcher checks only those statistics that appear to be out of line, he will introduce bias into his data.

10. The reader is referred to a document published by the American Psychological Association for advice on the form of his research report.

11. The introductory sections of the report should always contain a clear and concise statement of the purposes of the research. These sections should also outline the background of the problem and the theory on which it is based. They may sometimes form the basis of an article for journal publication.

12. The section of the report that describes the procedure should be sufficiently detailed to permit the reproduction of the study. This is not the easy matter that it may appear to be on the surface.

13. The results section of the report should contain statistical summaries and reductions of the data rather than the raw data. The conclusions drawn from the study should be clearly related to the hypotheses that were stated in the introductory sections.

14. The final section on implications should discuss the problem of "where do we go from here." The writer of such a section should avoid the temptation of throwing into it many wild ideas, but rather should it be a well-organized presentation of thoughts and concepts that emerge from the study.

15. Diagrams and tables should be as far as possible self-explanatory. Often they are appropriate substitutes for lengthy printed discussions.

16. The student should seek to publish at least an abbreviated account of his study so that the results are made available to the profession.

17. The American Documentation Institute represents a relatively new method of distributing technical information and one that may well supersede many of the functions formerly undertaken by journals.

Some Final Considerations | 16

In this book, an attempt has been made to familiarize the student with some of the methodologies that can be adapted to the purposes of educational research or that already have been adapted. The writer feels that this presentation may have left the impression that a knowledge of techniques placed at the command of a shrewd analytic intellect represents the essential ingredient of successful research. If this impression has been given, the author has been guilty of misrepresentation, so this final chapter has been written to impress the reader with the importance of other factors in a successful research enterprise. To some extent, this chapter must be speculative, for only a little is known concerning the personal and environmental conditions that are necessary for the production of creative research. Some of the available information will be given here, but an attempt will be made also to stimulate the reader's thinking concerning the problems of developing a creative research program.

Ability, Productivity, and Some Possible Reasons for Lack of Productivity

The conditions necessary for high-level creative work have been identified to only a limited degree. At present there is much specula-

tion concerning this matter—and a little research, much of which is stimulated by the worthy hope that it will be found that a democratic type of organization is most favorable to the creative process. Corroboration of that view will be hard to find, if it can be found at all without a radical reformulation of the problem. Certainly it will be many decades before sufficient information will have been acquired about this problem to advise the student concerning the environment he should seek out if he is to be creative in his research. Perhaps different students will have widely different requirements in this respect. However, there would be some agreement among graduate school faculties that certain students who seem to have all the intellectual skills necessary for creative talent often fail to produce original research, and that lack of productivity often has its roots in certain common causes that need to be given brief consideration here. The student who is aware of some of these conditions may find means of avoiding them.

First, there is the problem of excessive ambition. Every graduate school is familiar with the student who cannot find a problem worthy of his consideration. He sees his fellow graduate students as persons willing to study trivia that it is beneath his dignity to study. He is likely to spend time disparaging the efforts of his associates. It is most desirable not to be this type of student. Until a person has accomplished much in research, he is not in a position to criticize the simple-mindedness of his associates. Only accomplishment in research brings with it the right to criticize, except for the few who have established themselves as recognized critical reviewers. The phenomenon of the graduate student who is hypercritical of others is usually interpreted to be a symptom of defensiveness and of feelings of inadequacy. The student should be aware of this, and perhaps this awareness will help him avoid this error of the beginning researcher.

The foregoing remarks need to be qualified. Often the overambitious graduate student has the greatest potential as a creative research worker. The student who is content to study some commonplace problem that can be solved by routine methods is probably concerned mainly with achieving his doctoral degree and then leaving research forever. He has neither the ambition nor the creative talent for a career in research, which takes high ambition and a willingness to undertake a search for knowledge as a pursuit worth while in itself. Perhaps, also, a successful scientific career requires a certain imper-

meability to criticism and a tendency to pursue courses of action that others think are unproductive.

The writer also suspects that the graduate student who is most capable of generating novel research ideas is the one who is often least able to evaluate such ideas critically. This is hardly surprising, for critical abilities seem in themselves to inhibit the free flow of ideas. Those who have ideas may be expected to have many poor as well as many good ones, and they need help in sorting them into the one category or into the other. Such students need the help of faculty advisers who recognize worth-while and researchable ideas and who praise the student for them. Too often such students are met with a barrage of criticism directed toward their poor ideas and do not receive credit for their imaginative talent.

Another source of lack of productivity is where the researcher feels satisfied with his results only if they show high consistency with expectation. Part of this tendency is, no doubt, a fear of criticism. To present clear-cut results that can have only one possible interpretation places the scientist in a position beyond reproach and beyond criticism, but few experiments ever yield results of this kind. Most scientists in the behavioral area must be content with results that show some accordance with expectation but involve at least some small suspicion that the results could have arisen by chance. Difficulty in tolerating the ambiguity that such results provide has prevented many students from finishing well-conceived studies. Many completed studies remain unpublished for this same reason. This problem insofar as it arises in a graduate school of education is probably best handled through encouragement given by the faculty.

An interesting problem in this connection is posed by the work of Mendel, who, it will be remembered, counted the frequency with which smooth peas and wrinkled peas appeared in certain hybrids. Statisticians who have examined his results state that they manifest a much closer agreement with the frequencies that would be expected on the basis of his theory than are ordinarily encountered. Some have suspected that the Abbé may have seen that his results showed some departure from theoretical expectation and, fearing that his work might be rejected by the scientific body, made some adjustments in his data. Of course this is just a hypothesis, for we cannot possibly know with any certainty whether he did or did not tamper with his data. The moral to be learned is perhaps that even if Mendel did

adjust his data, the theory that it was designed to substantiate was rejected by one of the world's leading scientific societies, and it was thirty years before it became accepted.

Another difficulty that seems to arise in the case of some research workers and that chronically limits their productivity is failure to communicate with others during the early planning stages of the inquiry. Such communication, with the resulting exchange of ideas, seems to be an important factor in the development of the research worker. It permits both the critical review of research ideas as well as the development of these ideas to the point where they are researchable. True, there are some workers who communicate little with their associates and yet produce research of real value, but such workers are generally persons who would be considered to be of the highest capability, even by graduate school standards. Most graduate students require substantial interaction with their associates as a part of their education and as a part of the process of evolving a researchable problem.

Difficulties of communication often permeate the entire research process. Some students show evidence of being able to conduct a well-designed experiment even though they do not seem able to describe it to others at the time or to write a presentable account of what they have done at a later date. Perhaps the administrative solution to the difficulties of such students might be to team them with some of their more communicative associates who are less adept as experimentalists. Graduate schools may have some difficulty in accepting such a solution.

The Importance of the Social Atmosphere in Creative Work

Historians agree that creative work has been produced in quantity by mankind only at certain times in history. For reasons that are largely unknown, these periods of creativity have usually followed great wars. The social climate seems to be an essential factor in releasing creative talents, for one may assume that the available talent is the same from generation to generation.

The present writer is inclined to believe that much the same is true of the graduate student who wishes to engage in creative research. The existence of a proper social atmosphere is important for undertaking creative work at the highest level at which he is capable of operating.

It is doubtful whether graduate schools of education have been particularly successful in providing a social environment that is congenial to the development of creative talent. Schools of education, like most graduate schools, are designed for students who are willing to conform to a mass of rules and regulations, who take and pass required courses regardless of whether or not they consider them worth while, and who are willing to develop a dissertation about a problem of interest to the faculty. As Benjamin Bloom and his associates have shown at the University of Chicago, the college professor is interested in developing students similar to himself. There is nothing wrong with this—except for the fact that there is now considerable evidence that the creative person tends to be a nonconformist and is somewhat insensitive to the demands of the community in which he lives. Nonconformist students have always had their problems. One is reminded at this point that Oxford University was founded by a group of students from the University of Paris who did not like the way in which the latter institution was run.

Perhaps the nonconformist student who has never regarded himself as such may, by reading this, recognize the source of some of his difficulties and benefit by this insight. It is the writer's fond hope that some faculty members who read this may as a result develop softened attitudes toward students who find that bearing with the rules and regulations of a graduate school is distasteful. A sympathetic attitude toward the oddities that go along with creative talent would do much to generate in schools of education an atmosphere in which original research can thrive.

Another major difficulty in generating an atmosphere sympathetic toward research stems from the fact that few members of most school of education faculties engage in research and thus do not regard it as an activity about which they can speak with enthusiasm. The young research worker should probably be developed in an environment where research is pursued with an excitement that can almost be described as breathless anticipation. This is not the kind of atmosphere found in most schools of education, though it does exist in a few. This situation has not been remedied by the development of bureaus of educational research, few of which conduct work that might be described as original research or research as it is discussed in this book. Most of these bureaus devote their efforts to rendering advisory services or conducting surveys at a rather superficial level.

They have an important public relations role, and if they do not contribute to new knowledge, they do at least facilitate the dissemination of the old. They provide an atmosphere that probably encourages the development of the administrator, but none has acquired a reputation for developing research workers.

The limited scope of such bureaus and institutes has been unfortunate. Although they may have provided service to local communities, they have contributed but little to the development of educational research as a branch of endeavor in the behavioral sciences. The result is that scientific research related to educational problems has had no sponsors, except for the few in major graduate schools who conduct research for their own satisfaction. There is a real need for research institutes of education that conduct research of the type which has been stressed in this book. Such organizations would provide facilities for producing a whole generation of educational researchers dedicated to discovery in the field. Education would indeed then have a body of professional research workers, much as other fields have, and would not have to rely upon research traditions being carried on by a few overworked professors who are able to eke out a few hours a week of research work.

Creative Work Requires Prolonged and Sustained Effort

A major difficulty in the way of the graduate student producing research that might be called original is that this requires prolonged and enduring effort. One can be misled in this matter by the well-known fact that many important ideas have come to famous scientists at times when they were thinking about something else. This is a rather typical phenomenon among high-level scientists. However, what should not be missed in this matter is that these important ideas did not appear in individuals who had never sought to discover them. In every case, they appeared in individuals who had struggled long to find a solution, and it happened that the solution came at a moment when they were concerned with other matters. There seems no doubt that the creative person spends extended periods of great conscious effort when all of his energies are devoted to the solution of a problem; indeed his energies may be so completely channeled that he appears to have almost a detachment from the other aspects of life, even to the point where associates may consider him to be callous or

thoughtless. The absent-mindedness of the intellectual is just one symptom of this deep preoccupation. In this milieu of thoughtfulness the important ideas emerge, though often at a moment when the mind has turned to other things.

The graduate student, however talented he may be, is rarely able to devote his entire energies to the solution of a single problem. He has to be preoccupied with course work and with somewhat prosaic matters such as language requirements. It is probably for this reason that even the brilliant doctoral student rarely produces a brilliant doctoral dissertation. Conditions conducive to work of this quality are not provided for the graduate student.

What has been said up to this point fails to bring out the distinction between what may be termed creative research and routine investigation. Most master's theses and doctoral dissertations fall into the latter category. They are designed to test some fairly obvious hypothesis. They may be considered to develop, as it were, a territory that has already been well explored, but they are valuable, for it is the well-developed territory that yields riches. The graduate student as a researcher is a developer rather than an explorer. If he makes discoveries, they are minor if not a little prosaic. It is the explorer who makes the major discoveries, and for him too are reserved the special excitements of high adventure, the despairs of failure, and occasionally the thrill of genuine discovery. The developer's life is somewhat more tranquil and decidedly less venturesome. He knows a great deal about the territory in which he is operating. He knows with some certainty what will be the outcome of his labors. In contrast, the explorer is searching for something that he does not yet know except in the vaguest way, and like Columbus, he ultimately may not recognize what he finds.

Ghiselin (1954) has described this aspect of the creative process in a way that is particularly appropriate. He refers to the creative person as one who is struggling to realize the unrealized. What he wants to accomplish is something still outside of anything he can as yet conceive, but it is there in its vaguest outline at the periphery of consciousness. The creative person may spend a lifetime in struggling to find a medium through which he can realize this objective of peripheral awareness. Many never find it.

The characteristics of the research worker that have been noted

in this chapter present a problem in the matter of developing educational research. The problem is generated by the fact that many schools of education require their faculty to have had public school teaching experience. This tends to select persons who are unlikely to have much of the disposition that one might hope to see in the researcher. A person who has a real interest in theoretical problems or in such abstract matters as represent the very roots of research can hardly be expected to show a deep interest in the activities of classroom management as a possible lifetime pursuit. The outstanding research worker, with his ability to detach himself from his environment, might even be considered a poor risk as a teacher in the public schools. Schools of education should come to realize that the talents required by personnel who operate schools are probably quite different from those required in research.

The argument has often been put forward that in order to understand educational phenomena, it is necessary to have had the experience of teaching in a classroom. The argument is extremely persuasive, but its attractiveness is thoroughly superficial. Its inadequacy is seen when it is extended to other fields. Does the physicist have to experience a free fall before he is prepared to study free-falling bodies? In the case of the behavioral sciences, personal experience in a situation often produces the actual disadvantages of being unable to perceive the situation with any objectivity.

The Financial Support of Research

Any course in educational research should leave the student with a concept of its cost and some idea of how it is financed. Most of those who take courses in educational research do not expect to make it a lifetime pursuit, but many may become administrators who perhaps will have some responsibility for sponsoring research projects and for obtaining financial support for them. In order to fulfill this function of fund-raiser and expediter of research, it is necessary to have a concept of both the cost of research and the nature of available sources.

There is no way of estimating how much is spent annually at the present time on educational research. Much if not most of the money spent is channeled into activities related to the collection of administrative information. This activity is usually referred to locally as

"research," but would not fall within the meaning of the term as used in this book. There is no way of drawing a sharp line between administrative data-gathering and research, since there are all shades of intermediate activity. For this reason, one cannot obtain an estimate of current expenditures in educational research.

If the amount spent on educational research by large city school systems is examined, it is found to be relatively small, and annual budgets as large as $30,000 for research purposes are considered to be exceptionally large. In contrast, a major manufacturer of cameras will spend $100,000 on the development of a single mechanical innovation, simply because it is not possible to develop a new device for less money. Manufacturers just do not expect to derive benefits by investing small sums in research, partly because a substantial fraction of all research does not result in a useful product. On the other hand, boards of education commonly expect useful products to emerge from a relatively small investment, and they often derive little return from the money invested because the investment was too small.

Bureaus of educational research in universities rarely have much financial support and are often staffed by relieving a professor of a course so that he can then devote forty or more hours per week to the venture. Such bureaus, as has been pointed out, tend to be overloaded with work of a nonresearch character. There are just a few bureaus that have established full-time research professorships, with no strings attached as to what the incumbent is to do in the position. This is a direct support of research in a way that is sorely needed.

It has been possible during the last ten years for a few bureaus of educational research to obtain funds by participating in Department of Defense projects on personnel classification and training. Under this system, money is obtained for undertaking specific projects on which a bid has been submitted. Such projects vary in size, but most are between $20,000 and $40,000. These may seem to be large sums, but the cash received does not represent the cash that is actually invested in research personnel. The institution itself is usually entitled to a fraction of the money to cover its investment in buildings and their upkeep and management costs. A research contract of $20,000 might involve a deduction of $7000 for that type of overhead charges, leaving only $13,000 for the research itself. The latter sum might be sufficient for a high-level project supervisor for one-quarter time, and

a full-time research assistant and clerk. Thus $20,000 a year will provide for one full-time research worker. plus some help at both a lower and a higher level.

The value to an institution of developing a research program in this way has been a matter of controversy in the literature. Some institutions have felt that it has provided a real opportunity to develop a research program. Others have felt that it has not been a good method of expanding research because it means hiring temporary professional employees who cannot be given any guaranty of a permanent position. Perhaps the matter should also be regarded in a different light. Participation in defense projects, whatever other purposes it may serve, is a contribution to national defense.

Other government agencies have also given some support to educational research. Public Health Service research grants have been given for the support of projects related to matters of mental hygiene. The Department of Health, Education, and Welfare has recently expanded its activities to include the support of educational research. The National Science Foundation has shown interest in supporting educational research related to problems of training scientists, but the broad support of educational research is not anticipated from this source.

Both major and minor foundations have provided some support for educational research, and they are sources from which additional help may be sought. Proposals may be submitted to foundations, which will have them reviewed by competent persons in the field. Since World War II, there has been a tendency for foundations to provide institutions with rather large sums of money for long-range programs of research. This is in accordance with the concept that research is best carried out along programmatic lines. An immediate consequence of this is that the individual researcher has only limited opportunities for raising relatively small sums for his own projects.

The writer is a little unhappy about the policy that involves the donation of large sums to schools of education or bureaus of educational research to be spent within a relatively short period of time. This has been done, as in one case where an institution was given several million dollars for a five-year research project. On the surface this might seem to be a promising approach, for it does at least give recognition to the fact that productive research costs substantial

amounts of money. It does, however, neglect one important aspect of research that must be recognized if successful work is to be done: namely, that inquiry into a novel field is not likely to be accomplished successfully by persons who have not already spent much time thinking about related problems. The outstanding advances in the behavioral sciences seem to be made typically by scientists who have spent many years thinking about and working on a problem and who have developed a group of scientists with whom they work and who together have learned to speak a common language. It perhaps may have taken many years before the leading scientist in this group reached the point where his thoughts had become organized into a useful system.

After the last war the United States government attempted to buy research in the behavioral sciences on a grand scale. This action was taken on the assumption that creative ideas could be bought just like any other commodity. Such a procedure seems to have had some success in the production of technological developments, but the writer is unaware of any major scientific development that has been made in this way, and it would be surprising if this had happened.

Foundations expect a return for their money. For this reason, a foundation is likely to evaluate the person who submits an idea as carefully as it evaluates the idea itself. An idea may be a good one, but if it is submitted by a person who has little notion concerning how research should be conducted, there is no point in investing money in the project. For this reason, it is difficult for a person who does not have an established reputation to obtain a grant for original research. Some foundations do have small funds set aside for sponsoring those who are new in the research field, but such sponsorship is considered to be a long shot, with high chances that there will be no return on the investment.

Money for educational research is available, and perhaps one might say that it is available in quantity. What is lacking is a body of trained research workers with a broad scientific background who have the capacity for conducting programmatic research. The few who at present have these qualifications can utilize only a small fraction of the money that could be made available for educational research.

Appendix

An Example of a Theory from the Behavioral Sciences

It is difficult, if not impossible, to obtain from an educational science of behavior illustrations of a fairly complex theory of behavior stated with some degree of precision. A handful of simple theories involving one or two postulates can be found, but these are relatively few and far between. It is therefore necessary to turn to the field of psychology to provide a good illustration of a fairly well-stated theory that has some complexity. The illustration selected here is one developed by Ammons (1954), and it is shown in Table 4. It should be noted that all terms that have any special technical significance have been defined with some care. The theory as it is presented here includes only four postulates, but it has been derived from a more comprehensive theory containing twelve postulates.

Once he had stated his theory, Ammons went on to look for situations in which he could test some of the deductions from it. In one case, he found a situation that seemed to provide a rough test of deduction 4a. In an essay examination in one of his classes, some students were given four questions of which all were hard, while some were given four questions of which two were easy. It was hypothesized that when the questions

were all hard the "feelings" and drive of the students would be heightened and more error responses would occur. The data were found to be consistent with the hypothesis.

The reader should note that the theory may be described as a miniature system, and by this is meant that it is not a general theory of behavior but a theory of only a very limited segment of behavior. Ammons refers to it as a theory of error to indicate that it is concerned primarily with the prediction of error behavior. He has generally taken the stand that it is limited theories of behavior formulated with some precision that at present will serve best for the purpose of organizing knowledge of human behavior.

Much can be learned about the problems and difficulties of developing theories of behavior in education by examining this theory. First, note that the language of a scientific theory is rather different from the language ordinarily used for discussing behavior. The language refers to ideas and concepts that are quite different from those used generally in describing behavior. This is partly because the formulation of a theory requires that the phenomena with which the theory is concerned should be described as far as possible in terms of variables that can be measured. In the most advanced theories, such as those developed in the physical sciences, all terms are quantified and the postulates are represented by equations.

The advantages of stating a theory in terms of variables is obvious when we turn to a consideration of the next criticism that may be leveled against the theory presented. If the deductions from the theory are examined, it will be seen that they are not derived from the postulates in a manner that has any great rigor. They do follow from the postulates in a general sense, but they do not give the impression of following with absolute necessity.

The fact that most theories in the behavioral sciences are such that there is no rigorous method of making deductions from them results in much controversy over which theories are to be accepted and which are to be rejected. Indeed, it has happened on occasions, even though these are rare occasions, that certain evidence is accepted by one school of thought as substantiating a theory while the same evidence is accepted by another school of thought as rejecting it. Such conflicts are entirely the result of difficulties encountered in stating the postulates of the controversial theory in such a way that rigorous deductions made from them can be used to substantiate or reject the theory. As theories in the behavioral sciences become better and better formulated, this type of controversy will inevitably vanish.

Finally, it must be pointed out that the language used in describing the theory is not the language ordinarily used for describing behavior in daily conversation. The language of the theory, which is referred to as the data language, is usually a technical language of its own. The language of daily living does not lend itself well to the development of theory in the behavioral sciences.

TABLE 4

Theory of a limited aspect of behavior drawn up by Ammons in terms of postulates, definitions of the terms used in the postulates, and some deductions derived from the postulates.*

Definitions

Error: A response other than that appropriate to the motor set present, where this response is appropriate to other parts of the stimulus complex.

Response: Observable striated muscular behavior by the individual.

Motor set: Bodily orientation for the performance of a given behavior, inferred jointly from the instructions given by the experimenter or subject to himself and the physical orientation of the person. We can to some extent get at it by asking the subject what he intends or intended to do, or by setting up an objective criterion for determining whether or not the physical orientation would allow the performance of the task.

Appropriate response: The response which the individual says he intends or intended to make and for which he is physically oriented is the appropriate response to the motor set. Appropriate responses to other parts of the stimulus complex are those which would be most frequently made if those parts of the stimulus complex were dominant.

Stimulus complex: Various components which make up the stimulus such as stimuli from motor set, specific drive stimuli, and external stimuli. Any of these can be changed relatively independently, changing the stimulus complex.

Dominance of a component of the stimulus complex: A drive stimulus is more dominant as the drive becomes stronger. When the subject is asked to describe a situation, a particular stimulus component is dominant to the extent that it is mentioned earlier in his description.

Drive stimuli: Those stimuli characteristically noted by the human organ-

* Reprinted with permission from "'Errors': Theory and Measurement," by R.B. Ammons, in The Kentucky Symposium: *Learning Theory, Personality Theory, and Clinical Research*, copyright © 1954, John Wiley & Sons, Inc.

TABLE 4 (Continued)

ism in connection with hunger, thirst, sex frustration, fear, anxiety, etc. One could infer the presence of such stimuli in terms of strength of drive.

Definitions

External stimuli: Environmental energies which affect the receptors of the organism.

Strength of the response tendency: Latency of the response, physical strength of the response, and probability of the response occurring in the presence of or closely following the presence of a given stimulus complex.

Stimulus similarity: Stimulus complexes are similar to the degree that they contain similar components and are relatively less separated along the various discriminable continua.

Strength of drive: Might be the self-rating of the individual or might be inferred from the past history of the individual with respect to the time since drinking, time since eating, number of times a pleasant or unpleasant consequence has followed a particular stimulus complex, etc. Thus drive stimuli can be associated with primary or secondary drives as conceived of by Hull. Emotions are considered to be drives.

Reward: The satisfaction of some need, goal-object consumption, or avoidance of noxious stimulation.

Postulates

Postulate 1: To any stimulus component or complex, there are a number of possible responses. The strengths of the response tendencies differ. Thus there is present a "strength" hierarchy of responses to any given stimulus component or complex.

Postulate 2: The more similar a stimulus component or complex is to another given stimulus component or complex which has regularly elicited a response in the past, the stronger the response of this kind now elicited by the new stimulus.

Postulate 3: The stronger the drive, the stronger the response.

Postulate 4: The components of a given stimulus complex may in isolation elicit different responses. When the components are combined in the stimulus complex, the greater the dominance of a given component and the greater the strength of a given response tendency associated with it, the more likely the stimulus complex is to elicit this response.

Deductions

Deduction 3a: The more drive present, the less similar the external stim-

TABLE 4 (Continued)

ulus need be to the stimulus which in the past has regularly elicited a response, for it to be elicited with the same strength.

Deduction 4a: If a response has been regularly elicited under a low drive and is now elicited with a high drive of the same kind present, we will observe an increase in "errors," providing the strongest response tendencies to the motor set and the drive are different and that to the motor set is dominant.

Deduction 4b: If a response has been regularly elicited under one drive, and the drive is changed to another without altering the other stimulus components (especially motor set), there will be more errors, providing the appropriate dominant response to the drive-stimulus component from the original drive was the same as that to the motor set, but that to the new drive stimulus is different from that to the motor set, the motor set staying the same.

Deduction 4c: To the extent that a single stimulus component dominates the total stimulus complex, the successive responses given by an individual will be more similar to each other.

Strong emotion leads to stereotypy of responses, as does instruction-induced "motor set," and the "same" physical stimulation. In free association, problem areas will be talked about more frequently than other areas. In the case of errors, we find that certain kinds are quite frequent, i.e. certain types of slips of the tongue and certain kinds of accidents in the accident-prone person. These errors should indicate the life areas in which the person has problems and thus be of diagnostic value to the clinician.

Deduction 4d: Other stimulus conditions being approximately equal, if one arouses feeling about an error he should get real-life responses associated with a similar set, emotion, or drive more quickly than if no feeling is aroused.

Bibliography

Adler, Mortimer, J. "Liberalism and Liberal Education," *The Educational Record*, XX (1939), 422–424.

Alexander, Carter, and Arvid J. Burke. *How to Locate Educational Information and Data*. New York, Bureau of Publications, Teachers College, Columbia University, 1950.

Allport, Gordon W. *Personality*. New York, Henry Holt & Co., Inc., 1937.

American Psychological Association. *Technical Recommendations for Psychological Tests and Diagnostic Techniques*. Washington, D.C., American Psychological Association, 1954.

American Psychological Association Publication Manual, 1957 revision. Washington, D.C., American Psychological Association, 1957.

Ammons, Helen, and Arthur L. Irion. "A Note on the Ballard Reminiscence Phenomenon," *Journal of Experimental Psychology*, XLVIII (1954), 184–186.

Bales, R.F. *Interaction Process Analysis*. Cambridge, Mass., Addison Wesley Publishing Company, 1954.

———, and H. Gerbrands. "The Interaction Recorder," *Human Relations*, I (1948), 456–464.

Ballard, P.B. "Obliviscence and Reminiscence," *British Journal of Psychology, Monographs Supplement*, I, No. 2 (1913).

Berelson, Bernard. *Content Analysis*. Glencoe, Ill., The Free Press, 1952.

Bergmann, Gustav. *Philosophy of Science*. Madison, Wis., University of Wisconsin Press, 1957.

Bloom, Benjamin S. (ed.). *A Taxonomy of Educational Objectives*. New York, Longmans, Green & Co., Inc., 1956.

Boring, Edwin G. "Intelligence as the Tests Test It," *New Republic*, XXXV, No. 444 (1923), 35–37.

———. "The Nature and History of Experimental Control," *American Journal of Psychology*, LXVII (1954), 573–589.

Bowers, R.V. "Discussion and Analysis of the Problem of Validity." *American Sociological Review*, I (1936), 69–74.

Bruner, Jerome S., Jacqueline J. Goodnow, and George A. Austin. *A Study of Thinking*. New York, John Wiley & Sons, Inc., 1956.

Brunswick, Egon. "Representative Design and Probabilistic Theory in a Functional Psychology," *Psychological Review*, LXII (1955), 193–217.

———. *Systematic and Representative Design of Psychological Experiments*. ("University of California Syllabus Series," No. 304) Berkeley, 1947.

Bush, Robert R., and Frederick Mosteller. *Stochastic Models for Learning*. New York, John Wiley & Sons, Inc., 1955.

Campbell, Norman. *What is Science?* New York, Dover Publications, Inc., 1952.

Cantril, Hadley. *Gauging Public Opinion*. Princeton, N.J., Princeton University Press, 1947.

Carter, L., W. Haythorn, Beatrice Meirowitz, and J. Lanzetta. "A Note on a New Technique of Interaction Recording," *Journal of Abnormal and Social Psychology*, LXVI (1951), 258–260.

———, ———, ———, and ———. "The Relation of Categorizations and Ratings in the Observation of Group Behavior," *Human Relations*, IV (1951), 239–254.

Cochran, W.G., and Gertrude M. Cox. *Experimental Designs*. New York, John Wiley & Sons, Inc., 1950.

Coladarci, Arthur P. "Are Educational Researchers Prepared to Do Meaningful Research?" *California Journal of Educational Research*, V (1954), 3–6.

Conant, James B. *On Understanding Science*. New Haven, Conn., Yale University Press, 1946.

Coombs, C.H. "Theory and Methods of Social Measurement," in D. Katz and L. Festinger, (eds.). *Research Methods in the Behavioral Sciences*. New York, The Dryden Press, Inc., 1953.

Cornell, Francis G., Carl M. Lindvall, and Joe L. Saupe. *An Exploratory Measurement of Individualities of Schools and Classrooms.* University of Illinois, Bureau of Educational Research, L (1953). Bulletin 75.

Cronbach, Lee J. "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika,* XVI (1951), 297–334.

Dale, Edgar, and Jeanne Chall. "A Formula for Predicting Readability," *Educational Research Bulletin,* XXVII (1948), 11–20 and 37–54.

Davis, F.B. *Item Analysis Data: Their Interpretation, Computation, and Use in Test Construction.* Cambridge, Mass. Graduate School of Education, Harvard University, 1946.

Dewey, John. *How We Think.* Boston, D.C. Heath & Company, 1910.

Dollard, John, and Neal E. Miller. *Personality and Psychotherapy.* New York, McGraw-Hill Book Co., 1950.

Estes, W.K., S. Koch, K. MacCorquodale, P.E. Meehl, C.G. Mueller, W.N. Schoenfeld, and W.S. Verplanck. *Modern Learning Theory.* New York, Appleton-Century-Crofts, Inc., 1954.

Evaluative Criteria. Washington, D.C., Cooperative Study of Secondary School Standards, 1950.

Festinger, Leon, and Daniel Katz (eds.). *Research Methods in the Behavioral Sciences.* New York, The Dryden Press, Inc., 1953.

Flesch, Rudolf. *How to Test Readability.* New York, Harper & Brothers, 1951.

Freeman, Frank N. "Tests," *Psychological Bulletin,* XI (1914), 253–256.

French, Elizabeth. *Development of a Measure of Complex Motivation.* ("Research Report AFPTRC–TN–56–48.") Lackland Air Force Base, Texas, Air Force Personnel and Training Research Center, 1956. (Available from Armed Services Technical Information Agency Document Service Center, Dayton 2, Ohio.)

———. "Motivation as a Variable in Work-Partner Selection," *Journal of Abnormal and Social Psychology,* LIII (1956), 96–99.

Fries, C.C. *The Structure of English.* New York, Harcourt, Brace & Co., 1952.

Gardner, D.H. *Student Personnel Service.* Chicago, University of Chicago Press, 1936.

Getzels, J.W. "The Question-Answer Process: A Conceptualization and Some Derived Hypotheses for Empirical Examination," *Public Opinion Quarterly,* VIII (1954), 79–91.

Ghiselin, Brewster. *The Creative Process*. Berkeley, University of California Press, 1954.

Good, Carter V. *Dictionary of Education*. New York, McGraw-Hill Book Co., 1945.

Gray, W.S., and B.E. Leary. *What Makes a Book Readable*. Chicago, Chicago University Press, 1935.

Gringes, W.W. *Laboratory Instrumentation in Psychology*. Palo Alto, Cal., The National Press, 1954.

Guilford, J.P. *Psychometric Methods*, 2nd edn. New York, McGraw-Hill Book Co., Inc., 1954.

Gulliksen, H. "Intrinsic Validity," *American Psychologist*, V (1950), 511–517.

Hebb, Donald O. *The Organization of Behavior*. New York, John Wiley & Sons, Inc., 1949.

Heyns, Roger W., and Ronald Lippitt. "Systematic Observational Techniques," in Gardner Lindzey (ed.), *Handbook of Social Psychology*. Cambridge, Mass., Addison-Wesley Publishing Company, Inc., 1954.

Hull, C.L. *Principles of Behavior*. New York, Appleton-Century-Crofts, Inc., 1943.

Humphreys, Lloyd G. "Clinical Versus Actuarial Prediction," *Proceedings of the 1955 Invitational Conference on Testing Problems*, 129–135. Princeton, Educational Testing Service, 1956.

Jensen, Alfred C. "Determining Critical Requirements of Teachers," *Journal of Experimental Education*, XX (1951–1952), 79–85.

Kenny, Douglas T., and Sidney W. Bijou. "Ambiguity of Pictures and Extent of Personality Factors in Fantasy Responses," *Journal of Consulting Psychology*, XVII (1953), 283–288.

Kentucky Symposium. *Learning Theory, Personality Theory, and Clinical Research*. New York, John Wiley & Sons, Inc., 1954.

Lindquist, E.F. *Design and Analysis of Experiments in Psychology and Education*. Boston, Houghton Mifflin Co., 1953.

Loevinger, Jane, Goldine Gleser, and P.H. DuBois. "Maximizing the Discriminating Power of a Multiple Score Test," *Psychometrika*, XVIII (1953), 309–317.

Lorge, Irving. "Predicting Readability," *Teachers College Record*, XLV (1944), 404–419.

Lubin, Ardie, and Hobart G. Osborn. "A Theory of Pattern Analysis for the Prediction of a Quantitative Criterion." *Psychometrika,* XXII (1957), 63–73.

Manual of Accrediting Procedures. North Central Association of Colleges and Secondary Schools Commission on Institutions of Higher Education, 1934.

Marquis, Donald G. "Research Planning at the Frontiers of Science." *American Psychologist,* III (1948), 430–438.

McClelland, D.C., J.W. Atkinson, R.A. Clark, and E.L. Lowell. *The Achievement Motive.* New York, Appleton-Century-Crofts, Inc., 1953.

MacKinnon, D.W. Pp. 491–501 in H.A. Murray (ed.) *Explorations in Personality.* New York, Oxford University Press, 1938.

McPherson, Joseph H. *Predicting the Accuracy of Oral Reporting in Group Situations.* (*Research Bulletin,* 54–13.) Lackland Air Force Base, Texas, Air Force Personnel and Training Research Center, 1954.

McVey, William E. *Standards for the Accreditation of Secondary Schools.* Chicago, University of Chicago Press, 1942.

Medley, Donald M., Harold E. Mitzel, and Arthur N. Doi. *Analysis of Variance Models and Their Use in a Three-Way Design Without Replication.* (*Research Report,* No. 29.) New York, Division of Teacher Education, Board of Higher Education, 1955.

Meehl, Paul E. *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.* Minneapolis, University of Minnesota Press, 1954.

———. "Configural Scoring." *Journal of Consulting Psychology,* XIV (1950), 165–171.

———, and K. MacCorquodale. "On a Distinction Between Hypothetical Constructs and Intervening Variables." *Psychological Review,* LV (1948), 95–142.

———, and Albert Rosen. "Antecedent Probability and Efficiency of Psychometric Signs, Patterns, or Cutting Scores." *Psychological Bulletin,* LII (1955), 194–216.

Mill, J.S. *A System of Logic, Ratiocinative and Deductive.* New York, Longmans, Green & Co., Inc., 1930.

Miller, James G. "Toward a General Theory for the Behavior Sciences." *American Psychologist,* X (1955), 513–531.

Mitzel, Harold E., and William Rabinowitz. *Assessing Social-Emotional Climate in the Classroom by Withall's Technique.* (*Psychological*

Monographs, No. 368.) Washington, D.C., American Psychological Association, 1953.

Morsh, Joseph E. *Systematic Observation of Instructor Behavior*. Lackland Air Force Base, Texas, Air Force Personnel and Training Research Center, 1955.

———, George Burgess, and Paul N. Smith. *Student Achievement as a Measure of Instructor Effectiveness.* (Bulletin No. TN 55–12.) Lackland Air Force Base, Texas, Air Force Personnel and Training Research Center, 1955.

Mort, Paul R., and Walter C. Reusser. *Public School Finance*. New York, McGraw-Hill Book Co., 1951.

Mosier, C.I. "A Critical Examination of the Concepts of Face Validity," *Educational and Psychological Measurement*, VII (1947), 191–205.

Northrop, F.S.C. *The Logic of the Sciences and the Humanities*. New York, The Macmillan Company, 1948.

Olson, Willard C. *Child Development*. Boston, D.C. Heath & Company, 1949.

Parten, Mildred B. *Surveys, Polls, and Samples; Practical Procedures*. New York, Harper & Brothers, 1950.

Payne, S.L. "Interviewer Memory Faults," *Public Opinion Quarterly*, XIII (1949), 684–685.

Rabinowitz, W., and R.M.W. Travers. "Problems of Defining and Assessing Teacher Effectiveness, *Educational Theory*, III (1953), 212–219.

"Report of the Committee of the American Psychological Association on the Standardization of Procedure in Experimental Tests," *Psychological Monographs*, XIII (1910), No. 1, 107; No. 2, 53; No. 5, 85.

Revised Manual of Accrediting. Commission on Institutions of Higher Education, North Central Association of Schools and Colleges, 1941.

Ruger, Henry A., and Brencke Stoessinger. "On Growth Curves of Certain Characteristics in Man," *Annals of Eugenics*, II (1927), 76–110.

Russell, J.D., and F.W. Reeves. *Administration*. Chicago, University of Chicago Press, 1936.

———, and ———. *Finance*. Chicago, University of Chicago Press, 1935.

Seeman, Julius, and Nathaniel J. Raskin. "Research Perspectives in Client Centered Therapy," in O.H. Mowrer (ed.), *Theory and Research in Psychotherapy*. New York, The Ronald Press Company, 1951.

Skinner, B.F. *Behavior of Organisms.* New York, Appleton-Century-Crofts, Inc., 1938.

———. "A Case History in Scientific Method." *American Psychologist*, XI (1956), 221–233.

Soloman, R.L. "An Extension of Control Group Design," *Psychological Bulletin*, XLVI (1949), 137–150.

Spence, Kenneth W. *Behavior Theory and Auditioning.* New Haven, Yale University Press, 1956.

———. "Current Interpretations of Learning Data and Some Recent Developments in Stimulus-Response Theory," in Kentucky Symposium, *Learning Theory, Personality Theory, and Clinical Research.* New York, John Wiley & Sons, Inc., 1953.

Stephenson, William. *The Study of Behavior; A Technique and its Methodology.* Chicago, University of Chicago Press, 1953.

Stevens, S.S. "On the Theory of Scales of Measurement," *Science*, CIII (1946), 677–680.

Thomas, Frank, Elizabeth French, and R.M.W. Travers. "Variables Related to Problem Solving Effectiveness in Two Different Types of Problem Situation," *Proceedings of the 1955 National Research Council Symposium on Aviation Psychology.* Washington, D.C., National Research Council, 1956.

Thorndike, E.L. *Educational Psychology.* New York, Teachers College, Columbia University, 1913.

———. *Measurement of Intelligence.* New York, Teachers College, Columbia University, 1927.

———. *Theory of Mental and Social Measurement.* New York, Teachers College, Columbia University, 1913.

Tiedeman, David V., Joseph G. Bryan, and Philip J. Rulon. *The Utility of the Airman Classification Battery for Assignment of Airmen to Eight Air Force Specialties.* Cambridge, Mass., Educational Research Corporation, 1953.

Tomkin, S.S. *The Thematic Apperception Test; Theory and Technique of Interpretation.* New York, Grune & Stratton, Inc., 1947.

Travers, R.M.W. *Educational Measurement.* New York, The Macmillan Company, 1955.

———. *An Enquiry Into the Problem of Predicting Achievement.* (*Air Force Personnel and Training Research Center Bulletin*, 54–93.) Lackland Air Force Base, Texas, 1954.

———. "The Evaluation of the Outcomes of Guidance," *Educational and Psychological Measurement*, XVIII (1948), 325–333.

Tuddenham, Read D., and Margaret M. Snyder. *Physical Growth of*

California Boys and Girls from Birth to Eighteen Years. (*Publications in Child Development*, I, No. 2.) Berkeley, University of California Press, 1954.

Wallace, David. "A Case For and Against Mail Questionnaires," *Public Opinion Quarterly*, XVIII (1954), 40–52.

Waples, Douglas. *The Library*. Chicago, University of Chicago Press, 1936.

Wheeler, R.H. *The Science of Psychology*. New York, Thomas Y. Crowell Co., 1940.

Withall, John. "The Development of a Technique for the Measurement of Social Emotional Climate in the Classroom," *Journal of Experimental Education*, XVII (1949), 347–361.

Zimmer, Herbert. "Validity of Extrapolating Nonresponse Bias from Mail Questionnaire Follow-ups," *Journal of Applied Psychology*, XL (1956), 117–121.

Zook, George F., and M.E. Haggerty. *The Evaluation of Higher Institutions*. Chicago, University of Chicago Press, 1936.

———, and ———. *Principles of Accrediting Higher Institutions*. Chicago, University of Chicago Press, 1935.

Index of Names

Index of Subjects